12-2015

# MOBILITY AND ACTIVITY SPACE: UNDERSTANDING HUMAN DYNAMICS FROM MOBILE PHONE LOCATION DATA

Yang Xu
*University of Tennessee - Knoxville*, yxu30@vols.utk.edu

To the Graduate Council:

I am submitting herewith a dissertation written by Yang Xu entitled "MOBILITY AND ACTIVITY SPACE: UNDERSTANDING HUMAN DYNAMICS FROM MOBILE PHONE LOCATION DATA." I have examined the final electronic copy of this dissertation for form and content and recommend that it be accepted in partial fulfillment of the requirements for the degree of Doctor of Philosophy, with a major in Geography.

Shih-Lung Shaw, Major Professor

We have read this dissertation and recommend its acceptance:

Bruce Ralston, Hyun Kim, Lee Han

Accepted for the Council:

Carolyn R. Hodges

Vice Provost and Dean of the Graduate School

(Original signatures are on file with official student records.)

# MOBILITY AND ACTIVITY SPACE: UNDERSTANDING HUMAN DYNAMICS FROM MOBILE PHONE LOCATION DATA

A Dissertation Presented for the

Doctor of Philosophy

Degree

The University of Tennessee, Knoxville

Yang Xu

December 2015

# DEDICATION

I dedicate this dissertation work to my parents, Hui Xu and Jun Zheng, who encouraged me to pursue happiness and meanings in my life. Their love and support have given me great strength to overcome the challenges in my pursuit of the Ph.D. degree.

I dedicate this dissertation work to my wife, Wenting Pi, who taught me that the best love is to accompany.

I also dedicate this dissertation to my advisor, Dr. Shih-Lung Shaw, for his guidance, patience and encouragement through the past four years.

# ACKNOWLEDGEMENTS

First and foremost, I want to express my gratitude to my advisor, Dr. Shih-Lung Shaw. He is such a person who cares about me, my research, and my future career. This dissertation work could not have been completed without his inspiration and guidance. Through the past four years, he has spent so much effort and energy teaching me the ethics of conducting research. He always encourages me to think independently, creatively and critically. He has been so patient to nurture me through the various research challenges that I've encountered. Everything that I've learned from him will benefit me for the rest of my life.

My sincere appreciation goes to Dr. Bruce Ralston, who has been a great help to me on both my course work and dissertation research. I've gained numerous skills from him through his courses in GIS, transportation and computer programming. He offered me valuable suggestions to the writing of this dissertation.

I want to thank my other committee members, Dr. Hyun Kim and Dr. Lee Han, for spending time with me discussing the details in my dissertation. They offered me great suggestions on how this dissertation could be extended for my future research within a broader scope of GIS and transportation.

My special thanks go to Dr. Qingquan Li, who served as my supervisor when I was a master student in China. He led me to the wonderful world of GIS and encouraged me to pursue my Ph.D. degree in the United States. I would never be able to reach this point without his kind support.

I would like to thank the professors, graduate students and staff in the Geography Department. Their kind help and support make my stay in this department a wonderful experience. Finally, I must thank all of my friends at Knoxville who have always been with me in good times and bad. Your presence has become one of the best memories in my life.

# ABSTRACT

Studying human mobility patterns and people's use of space has been a major focus in geographic research for ages. Recent advancements of location-aware technologies have produced large collections of individual tracking datasets. Mobile phone location data, as one of the many emerging data sources, provide new opportunities to understand how people move around at a relatively low cost and unprecedented scale. However, the increasing data volume, issue of data sparsity, and lack of supplementary information introduce additional challenges when such data are used for human behavioral research. Effective analytical methods are needed to meet the challenges to gain an improved understanding of individual mobility and collective behavioral patterns.

This dissertation proposes several approaches for analyzing two types of mobile phone location data (Call Detail Records and Actively Tracked Mobile Phone Location Data) to uncover important characteristics of human mobility patterns and activity spaces. First, it introduces a home-based approach to understanding the spatial extent of individual activity space and the geographic patterns of aggregate activity space characteristics. Second, this study proposes an analytical framework which is capable of examining multiple determinants of individual activity space simultaneously. Third, the study introduces an anchor-point based trajectory segmentation method to uncover potential demand of bicycle trips in a city.

The major contributions of this dissertation include: (1) introducing an activity space measure that can be used to evaluate how individuals use urban space around where they live; (2) proposing an analytical framework with three individual mobility indicators that can be used to summarize and compare human activity spaces systematically across different population groups

or geographic regions; (3) developing analytical methods for uncovering the spatiotemporal dynamics of travel demand that can be potentially served by bicycles in a city, and providing suggestions for the locations and daily operation of bike sharing stations.

# TABLE OF CONTENTS

# LIST OF TABLES

# LIST OF FIGURES

# CHAPTER 1

# INTRODUCTION

## 1.1 Research Background and Research Questions

The rapid development of information and communication technologies (ICT) has brought remarkable changes to our society. The advent of mobile phone has transformed the way we interact with the outside world. According to the ICT statistics provided by the International Telecommunication Union (2015), there will be more than 7 billion mobile cellular subscriptions worldwide by the end of 2015, corresponding to a penetration rate of 97%. Mobile phone has become essential for people in many parts of the world to communicate with others and perform various tasks in their daily lives. The pervasiveness of mobile phone makes it a promising way to understand the dynamics of people's communication and activity patterns. Due to the proliferation of location-aware technologies such as Global Positioning System (GPS), Wi-Fi-based positioning system (WPS) and Bluetooth, we are now able to use mobile phones to locate people in space and time without posing much additional burden on them. If such data can be leveraged to study human dynamics in terms of their mobility patterns and use of space, it would benefit a variety of applications such as urban planning, public transport design, public health, emergency response and location-based services.

Traditional approaches that capture human activity patterns rely heavily on surveys, which are costly and time consuming (Levinson and Kumar 1995; Schönfelder and Axhausen 2003). Moreover, the amount of information that can be collected is largely restricted by the available human and financial resources. Mobile phone location data can be automatically collected by cellular carriers. The enormous quantity of data provides us with new opportunities to investigate various aspects of human dynamics from individual mobility patterns to collective behavior of the masses (Candia et al. 2008). For example, Call Detail Records (CDRs) collected by cellular companies as a byproduct of their daily operations contain information about when and where instances of cellphone calls and text messages occurred. Due to a large spatial, temporal and population coverage of CDRs, this type of data has been widely used in studies to investigate human mobility patterns and various aspects of urban dynamics (Ahas et al. 2007; Gonzalez, Hidalgo and Barabasi 2008; Ahas et al. 2010; Phithakkitnukoon et al. 2010; Cho, Myers and Leskovec 2011; Yuan, Raubal and Liu 2012; Becker et al. 2013; Silm and Ahas 2014; Alexander et al. 2015; Dong et al. 2015). Erlang data, another type of mobile phone data that measure person-hours of phone usage, have also been used in studies to better understand population distributions and daily rhythms of urban mobility (Ratti et al. 2006; Ratti et al. 2007; Reades, Calabrese and Ratti 2009; Sevtsuk and Ratti 2010). In recent years, there have been some new ways to collect mobile phone location data for social and behavioral research. For example, the Nokia Research Center conducted the Lausanne Data Collection Campaign (LDCC) in 2012 to collect "quasi-continuous" measurements of phone users' daily activities unobtrusively (Laurila et al. 2012). LDCC collected longitudinal mobile phone location data for 185 participants over a year using various types of technologies such as GPS, WLAN, Bluetooth, motion detection and so forth. The dataset collected by the LDCC is considered as an example of Actively Tracked Mobile Phone Location Data. This advanced data collection method makes it possible to capture individual location information more actively and produces datasets with finer temporal granularities. In general, the location information of Actively Tracked Mobile Phone Location Data id recorded when particular mobile phone "events" (e.g., periodic location update and handover of mobile phone signal) or communication activities (e.g., phone calls, text messages and usage of cellular data) occur. Table 1.1 gives an overview of the three types of mobile phone location data that have been used in previous studies.

Table 1.1 An overview of three major types of mobile phone location data used in human behavioral research and urban studies.

| Data Type | Characteristics |
|---|---|
| Call Detail Records (CDRs) | Location information is collected when a mobile phone user places or receives a phone call or text message. This type of data is also known as “passive” mobile phone location data. |
| Erlang Data | A measure of person-hours of cellphone usage. For example, 1 Erlang represents one person talking on the phone for an hour or two people talking for half an hour each. |
| Actively Tracked Mobile Phone Location Data | Data are collected using mixed types of positioning technologies such as GPS, WLAN and Bluetooth (Laurila et al. 2012). Location information is recorded when particular mobile phone “events” (e.g., periodic location update and handover of mobile phone signal) or communication activities (e.g., phone calls, text messages and usage of cellular data) occur. |

Understanding human mobility patterns and people’s use of space have always been an important topic in geography and transportation research. How people move around in their daily lives has great implications for the built environment and transportation systems. The unique characteristics of mobile phone location data make it an appealing resource for understanding the whereabouts of people in space and time. First, mobile phone location data can be automatically collected through our cellphones. The data collection method poses little burden on individual participants. The information collected is in digital format and can be readily processed by computers. The ease of data collection mechanism makes it possible to gather mobile phone location data frequently and economically. Moreover, mobile phone location data is able to reach a scale with tremendous spatial, temporal and population coverage, which opens new opportunities for large-scale human behavioral analysis. Despite of these advantages, mobile phone location data fall short of collecting individual socioeconomic and demographic information for travel behavior studies and policy analysis. Individuals' location records do not explicitly reflect their travel purposes. Second, mobile phone location data (e.g., CDRs) usually contain location information associated with particular kinds of human activities (e.g., mobile phone calls, text messages and cellular data usage), which reflects a partial aspect of individual daily activity patterns. Furthermore, the positioning accuracy of mobile phone data is largely affected by the density of cellphone towers in a city as well as load balancing and signal noise. Special considerations are needed to tackle spatial uncertainty in order to extract meaningful information about human activity patterns. Table 1.2 summarizes the relative strengths and limitations of using mobile phone location data in human mobility research.

Mobile phone location data provide both opportunities and challenges to the study of human mobility patterns in geography. It is thus very important to understand what kinds of research questions can be answered by this newly emerged data source as well as what kinds of human mobility patterns can be uncovered. In geography, one important concept related to human mobility patterns is human activity space. Activity space denotes the daily environment that an individual is using for his/her daily activities (Golledge and Stimson 1997). There are several related concepts such as awareness space (Brown and Moore 1970), action space (Horton and Reynolds 1971), perceptual space (Relph 1976) and mental maps (Lynch 1960). In general, an individual's activity space is usually conceptualized as the locations that have been visited as well as the travels among these locations (Schönfelder and Axhausen 2003). Human activities demonstrate a high degree of spatial and temporal regularity (Gonzalez, Hidalgo and Barabasi

2008; Song et al. 2010; de Montjoye et al. 2013). People perform their daily routines mainly at a few activity locations such as home, school, workplace, supermarkets, favorite restaurants and so forth. These locations are often considered as anchor points (Cullen and Godson 1975; Golledge and Stimson 1997) of individual activity space. Although mobile phone location data provide only snapshots of individual footprints in space and time, these individual footprints tend to concentrate at people's major activity locations. Thus, it is considered as a good data source for uncovering individual activity anchor points as well as the movement patterns among them (Ahas 2010; Montoliu and Gatica-Perez 2010). Such information embedded in mobile phone location data enables us to better understand people's use of space and its relationship with the built environment.

Table 1.2 Relative strengths and limitations of using mobile phone location data in human mobility research.

| **Strengths** | **Limitations** |
|---|---|
| • Data collection poses little burden on individual participants.<br>• The information collected is in digital format and can be readily processed by computers.<br>• Data can be collected more frequently with lower cost as compared to traditional travel surveys.<br>• Data usually have a large spatial, temporal and population coverage, which facilitates large-scale human behavioral research. | • Data fall short of collecting individual socioeconomic and demographic information for travel behavior and policy analysis.<br>• Data usually contain location information tied to particular kinds of human activities (e.g., mobile phone calls, text messages and cellular data usage), which reflects a partial aspect of individual daily activity patterns.<br>• Position accuracy is affected by cellphone tower densities, load balancing and signal noise. |

In order to describe people's daily activity patterns, considerable focus has been placed on measuring the size, geometry and structure of human activity space. For example, Dijst (1999) proposed an action space model with several discriminant functions to measure the size and shape (e.g., line, circle and ellipse) of individual activity space in order to better understand the activity patterns of two-earner families in two Dutch communities. Schönfelder and Axhausen (2003, 2004) used three measurement approaches (confidence ellipse, kernel density estimates and minimum spanning tree) to describe the structures of people's activity spaces based on travel surveys and GPS tracking datasets. Buliung and Kanaroglou (2006) developed a GIS toolkit to support exploration of household travel behavior using several activity space measures (e.g., standard deviational ellipse and standard distance). Similarly, Sherman et al. (2005) applied several activity space measures (standard deviational ellipse, road network buffer and travel time polygon) to examine individual accessibility to health care opportunities. They also discussed the relative strengths and limitations of each activity space measure in representing individual health accessibility. The measures mentioned above can effectively describe people's use of space from different perspectives. However, measures like confidence ellipse, standard deviational ellipse and standard distance often consider the arithmetic mean of individual activities as the center of activity space when measuring the spatial dispersion of visited locations. As suggested by Schönfelder and Axhausen (2003), an individual's home location should be treated as an important anchor point when analyzing how people move around in their daily lives. Understanding how people move around their home locations could yield insights into people's daily activity patterns and their interactions with the built environment. Moreover, understanding of people's use of space around their home locations could potentially benefit urban planning and public transport design for different neighborhoods in a city.

Although CDR data can be sparse in both space and time, they serve as a valuable data source for uncovering individual activity anchor points such as home locations, and people's use of space around these locations. Hence, the first set of research questions that this dissertation aims to answer is as follows:

- How do people's daily activities take place around their home location? How can Call Detail Records (CDRs) be used to estimate important individual activity anchor points (e.g., home anchor point) and to quantify the size of individual activity space centered at the home anchor point? Do people who live in different parts of a city exhibit different characteristics of activity space size? What are the places where people share similar characteristics of activity space size?

As pointed out by Miller and Goodchild (2014), "the context for geographic research has shifted from a data-scarce to a data-rich environment" (p. 449). Mobile phone location data and other newly-emerged data sources (e.g., social media data) provide rich and dynamic geo-referenced information for studies of human mobility patterns. New analytical methods are needed in the era of data-driven geography (Miller and Goodchild 2014) to respond to the challenges of increasing data volume, velocity and variety. In recent years, various measures such as radius of gyration (Song, Blumn and Barabási 2010), activity anchor point (Phithakkitnukoon et al. 2010; Cho, Myers and Leskovec 2011) and daily activity range (Isaacman et al. 2010; Becker et al. 2013) have been suggested to reflect major characteristics of individual activity space. However, several research challenges remain. For example, many previous studies examined important determinants of human activity space independently. It remains unclear how these determinants of individual activity space are related to each other. Although clustering methods have been applied to multi-dimensional analysis of human mobility such as identifying similar location sequence (Li et al. 2008), commuting flexibility (Shen, Kwan and Chai 2013) and spatiotemporal activity patterns (Chen et al. 2011), it sometimes can be difficult to interpret the major characteristics of each population group derived from the clustering algorithms. Moreover, these clustering methods (e.g., hierarchical clustering) are computationally intensive and often perform inefficiently over very large datasets. It is thus important to develop useful methods that summarize the major characteristics of individual activity space, and also reveal the relationships among different activity space determinants.

Spatial extent, frequent locations and movements are three major determinants of individual activity space (Golldge and Stimson 1997). These determinants provide much needed information for the understanding of human mobility patterns, and the similarities and disparities among different population groups or geographic regions. However, few studies have tried to analyze these activity space determinants simultaneously using mobile phone location data. To fill the research gap, this dissertation aims to answer the second set of research questions using Actively Tracked Mobile Phone Location Data collected in two major cities in China:

- What are the major characteristics of individual activity space in terms of spatial extent, frequent locations and movements in a city? How are these important determinants associated with each other in an individual's activity space? What are the major differences and similarities of human activity spaces across different cities? How do aggregate activity space patterns (e.g., daily activity range and movement patterns) change over space and time in a city?

Recently, there have been several studies which used mobile phone location data for travel demand forecast. Due to the capabilities of revealing important individual activity anchor points and movement patterns, such datasets can be very useful to study individual trip making, especially for trips that are tied to people's key activity locations (e.g., home and workplace). For example, Iqbal et al. (2014) used CDRs in Dhaka, Bangladesh to generate tower-to-tower transient OD matrices, which were then associated with traffic network and converted to node-to-node transient OD matrices. The estimation result was validated using traffic count data. Similarly, Alexander et al. (2015) used CDRs in Boston metropolitan area over a period of two months to estimate origin-destination trips by purpose (e.g., home-based work trips, home-based other trips and non-home based trips). The estimation proves to be relatively consistent with various national and local travel surveys. Dong et al. (2015) used mobile phone location data to suggest traffic zone division in urban areas to assist travel demand forecasting. Wang et al. (2012) used mobile phone location data collected in San Francisco Bay area and Boston area to evaluate urban road usage patterns. It is clear that mobile phone location data can be leveraged to better understand people's travel demand related to different types of human activities. Individual activity anchor points and their movement patterns serve as critical information in travel demand analysis.

Travel demand forecast could benefit transportation planning in a variety of ways such as providing suggestions to the design of pedestrian walkways or locations of appropriate cycling infrastructures. In recent years, a growing number of cities are promoting bicycle use to mitigate urban problems related to public health, traffic congestion, energy consumption and air pollution. In order to establish a good public bicycle program, it is important to know where the demands are and when they occur. Large-scale mobile phone location datasets have introduced new opportunities for researchers and planners to assess travel demand related to different transportation modes in a city. However, there has been limited research on how to estimate potential demand of bicycle trips from this newly emerged data source to facilitate planning and daily operation of cycling facilities such as bike sharing stations in a city. Hence, this dissertation aims to answer the third set of research questions using Actively Tracked Mobile Phone Location Data:

- How to extract meaningful trip chain segments from individual cellphone trajectories to better understand individual travel behavior and trip making? How to estimate potential demand of bicycle trips in a city from these extracted trip chain segments? How do such demands change over space and time in a city? If we want to locate a certain number of bike sharing stations in a city, where should we put them to best accommodate people's travel needs? What are the relationships between the incoming and outgoing trips at the suggested bike sharing stations at different times of a day, and what are their implications for the planning and daily operation of these bike sharing stations?

This study proposes several approaches for processing and analyzing two types of mobile phone location data (CDRs and Actively Tracked Mobile Phone Location Data) to address the above three sets of research questions. The dissertation research aims to provide some insights into intrinsic characteristics of human activity space, and discuss their relationships with the built environment and implications in urban design and transportation planning.

## 1.2 Organization of the Dissertation

This dissertation is organized in a manuscript format that includes three manuscripts targeted for different peer-reviewed journals.

Chapter 2 proposes a home-based approach to studying human activity space using a 13-day CDR dataset which covers more than 1 million individuals. The main objective of this manuscript is to better understand how people's daily activities take place around their home locations. The manuscript first introduces how each individual's "home" anchor point is estimated from CDRs. A modified standard distance (${S'}_D$) is proposed to measure the spread of each individual's activity space centered at this "home" anchor point. Aggregate activity space patterns are then derived at the mobile phone tower level to describe the distribution of ${S'}_D$ for individuals who share the same "home" anchor point. Finally, a hierarchical clustering algorithm is performed to find areas with similar aggregate activity space patterns. This manuscript aims to provide insights into people's use of space relative to their home anchor point, as well as the geographic disparities of human activity space in a city.

Chapter 3 uses two actively tracked mobile phone location datasets that cover a weekday to characterize people's use of space in two major cities in China. Three mobility indicators (daily activity range, number of activity anchor points, and frequency of movements) are introduced to represent the major determinants of individual activity space. By applying association rules in data mining, this manuscript analyzes how these three indicators are related to each other in each individual's activity space. The manuscript further examines the spatial and temporal variations of aggregate human mobility patterns in these two cities. This study aims at providing a multi-dimensional view of individual activity space and facilitating the comparison of human activity spaces across different cities.

Chapter 4 uses an actively tracked mobile phone location dataset to uncover potential demand of bicycle trips in a city. The study introduces an anchor-point based trajectory segmentation method to partition the cellphone trajectories into meaningful trip chain segments. By selecting trip chain segments that can be potentially served by cycling, two indicators ($inflow_p$ and $outflow_p$) are generated at the cellphone tower level to reflect the potential demand of incoming and outgoing bicycle trips at different places in a city and times of a day. A location-allocation model is performed to suggest facility locations of bike sharing stations based on the total demand generated at each cellphone tower. Two measures (accessibility and dynamic relationships between incoming and outgoing trips) are introduced to further understand the characteristics of the bike stations once their locations are derived. The research aims to develop analytical methods to better understand human travel behavior using mobile phone location data, and to provide suggestions to planning and daily operation of bike sharing stations in a city.

Chapter 5 summarizes the major contributions of this dissertation research to studies of human mobility patterns and activity space in geography, as well as implications to urban and transportation planning. This dissertation ends with discussions of future research directions.

## References

Ahas, R., A. Aasa, Ü. Mark, T. Pae, and A. Kull. 2007. Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management* 28 (3):898-910.

Ahas, R., A. Aasa, S. Silm, and M. Tiru. 2010. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies* 18 (1):45-54.

Ahas, R., S. Silm, O. Järv, E. Saluveer, and M. Tiru. 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology* 17 (1):3-27.

Alexander, L., S. Jiang, M. Murga, and M. C. González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58: 240-250.

Becker, R., R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. 2013. Human mobility characterization from cellular network data. *Communications of the ACM* 56 (1):74-82.

Brown, L. A., and E. G. Moore. 1970. The intra-urban migration process: a perspective. *Geografiska Annaler. Series B, Human Geography* 52 (1):1-13.

Buliung, R. N., and P. S. Kanaroglou. 2006. A GIS toolkit for exploring geographies of household activity/travel behavior. *Journal of Transport Geography* 14 (1):35-51.

Candia, J., M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41 (22):224015.

Chen, J., S.-L. Shaw, H. Yu, F. Lu, Y. Chai, and Q. Jia. 2011. Exploratory data analysis of activity diary data: a space–time GIS approach. *Journal of Transport Geography* 19 (3):394-404.

Cho, E., S. A. Myers, and J. Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1082-1090.

Cullen, I., and V. Godson. 1975. Urban networks: the structure of activity patterns. *Progress in planning* 4:1-96.

Dijst, M. 1999. Two-earner families and their action spaces: A case study of two Dutch communities. *GeoJournal* 48 (3):195-206.

Dong, H., M. Wu, X. Ding, L. Chu, L. Jia, Y. Qin, and X. Zhou. 2015. Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C: Emerging Technologies* 58: 278-291.

de Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports* 3(1376): 1-5.

Golledge, R., and R. Stimson. 1997. *Spatial Behavior*. Guilford, London.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196):779-782.

Horton, F. E., and D. R. Reynolds. 1971. Effects of urban spatial structure on individual behavior.*Economic Geography*:36-48.

Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. 2010. A tale of two cities. In Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications. pp. 19-24.

Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González. 2014. Development of origin–destination matrices using mobile phone call data. *Transportation Research Part C: Emerging Technologies* 40:63-74.

ITU. ICT Facts and Figures – The world in 2015. http://www.itu.int/en/ITU-D/Statistics/Documents/facts/ICTFactsFigures2015.pdf (last accessed on Sep 3, 2015)

Laurila, J. K., D. Gatica-Perez, I. Aad, O. Bornet, T.-M.-T. Do, O. Dousse, J. Eberle, and M. Miettinen. 2012. The mobile data challenge: Big data for mobile computing research. In Proceedings of Mobile Phone Data Challenge by Nokia Workshop. Colocated with Pervasive' 12.

Levinson, D., and A. Kumar. 1995. Activity, travel, and the allocation of time. *Journal of the American Planning Association* 61 (4):458-470.

Li, Q., Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. 2008. Mining user similarity based on location history. In Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems. Article No. 34.

Lynch, K. 1960. *The image of the city*: MIT press.

Miller, H. J., and M. F. Goodchild. 2014. Data-driven geography. *GeoJournal*:1-13.

Montoliu, R., and D. Gatica-Perez. 2010. Discovering human places of interest from multimodal mobile phone data. In Proceedings of the 9th International Conference on Mobile and Ubiquitous Multimedia. Article No. 12.

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. *Human Behavior Understanding*, 14-25: Springer.

Ratti, C., A. Sevtsuk, S. Huang, and R. Pailer. 2007. *Mobile landscapes: Graz in real time*: Springer.

Ratti, C., S. Williams, D. Frenchman, and R. Pulselli. 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33 (5):727.

Reades, J., F. Calabrese, and C. Ratti. 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design* 36 (5):824-836.

Relph, E. 1976. *Place and placelessness*: Pion London.

Schönfelder, S., and K. W. Axhausen. 2003. Activity spaces: measures of social exclusion? *Transport Policy* 10 (4):273-286.

Schönfelder, S., and K. Axhausen. 2004. Structure and innovation of human activity spaces. *Arbeitsberichte Verkehrs-und Raumplanung* 258: 1-40.

Sevtsuk, A., and C. Ratti. 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology* 17 (1):41-60.

Shen, Y., M.-P. Kwan, and Y. Chai. 2013. Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. *Journal of Transport Geography* 32:1-11.

Sherman, J. E., J. Spencer, J. S. Preisser, W. M. Gesler, and T. A. Arcury. 2005. A suite of methods for representing activity space in a healthcare accessibility study. *International Journal of Health Geographics* 4 (24):1-21.

Silm, S., and R. Ahas. 2014. Ethnic differences in activity spaces: A study of out-of-home nonemployment activities with mobile phone data. *Annals of the Association of American Geographers* 104 (3):542-559.

Song, C., Z. Qu, N. Blumm, and A.-L. Barabási. 2010. Limits of predictability in human mobility. *Science* 327 (5968):1018-1021.

Wang, P., T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González. 2012. Understanding road usage patterns in urban areas. *Scientific reports* 2 (1001):1-6.

Yuan, Y., M. Raubal, and Y. Liu. 2012. Correlating mobile phone usage and travel behavior–A case study of Harbin, China. *Computers, Environment and Urban Systems* 36 (2):118-130.

# CHAPTER 2

# UNDERSTANDING AGGREGATE HUMAN MOBILITY PATTERNS USING PASSIVE MOBILE PHONE LOCATION DATA – A HOME-BASED APPROACH

This chapter is a slightly modified version of a paper published in a peer-reviewed journal: Xu, Y., S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li. 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. *Transportation* 42(4):625-646. The paper was included in the journal's special issue on "Emerging, Passively Generated Datasets for Travel Behavior and Policy Analysis" in 2015. The use of "we" in this chapter refers to all the co-authors and myself. As the first author, I processed the data, performed the analysis and wrote the manuscript.

**Abstract:** Advancements of information, communication and location-aware technologies have made collections of various passively generated datasets possible. These datasets provide new opportunities to understand human mobility patterns at a low cost and large scale. This study presents a home-based approach to understanding human mobility patterns based on a large mobile phone location dataset from Shenzhen, China. First, we estimate each individual's "home" anchor point, and a modified standard distance ($S_D{}'$) is proposed to measure the spread of each individual's activity space centered at this "home" anchor point. We then derive aggregate mobility patterns at mobile phone tower level to describe the distance distribution of $S_D{}'$ for people who share the same "home" anchor point. A hierarchical clustering algorithm is performed and the spatial distributions of the derived clusters are analyzed to highlight areas with similar aggregate human mobility patterns. The results suggest that 43% of the population sample travelled within a short distance ($S_D{}' \leq 1km$) during the 13-day study period while 23.9% of them were associated with a large activity space ($S_D{}' \geq 5km$). The geographical differences of people's mobility patterns in Shenzhen are evident. Areas with a large proportion of people who have a small activity space mainly locate in the northern part of Shenzhen such as Baoan and Longgang districts. In the southern part where the economy is highly developed, the percentage of people with a larger activity space is higher in general. The findings could offer useful implications on policy and decision making. The proposed approach can also be used in other studies involving similar spatiotemporal datasets for travel behavior and policy analysis.

**Keywords**: Passive mobile phone location data, human mobility, activity space, home-based approach.

## 2.1 Introduction

Understanding human mobility patterns has been an important topic in transportation research. Traditional studies rely heavily on travel surveys and questionnaires to analyze people's movements and activities (Levinson and Kumar 1995; Dijst 1999; Axhausen et al. 2002; Schönfelder and Axhausen 2003). Such datasets usually consist of detailed information of respondents as well as the trips they made that can facilitate analysis of human travel and activity patterns. However, collection of such datasets can be costly and time-consuming and the sample size is constrained by the human and financial resources available. Recent advancements in location-aware technologies have provided many passively generated datasets for understanding the whereabouts of people in space and time. These datasets refer to a set of data that are passively generated only when particular kinds of human activities occur (e.g., cell phone usage, credit card transaction and check-in of social networks). Mobile phone location data, as one example of such datasets, "enables the study of human mobility at low cost and on an unprecedented scale" (Becker et al. 2013, p. 74). These large-scale anonymized datasets have introduced new opportunities to investigate various aspects of human dynamics from individual activity patterns to collective behavior of the masses (Candia et al. 2008).

In recent years, there have been many studies which used mobile phone location data to understand people's travel and activity patterns as well as the implications to transportation planning, urban design and social dynamics (Ahas et al. 2007; Candia et al. 2008; Gonzalez, Hidalgo and Barabasi 2008; Phithakkitnukoon 2010; Song et al. 2010; Cho, Myers and Leskovec 2011; Yuan, Raubal and Liu 2012; Becker et al. 2013). Most of them utilized passive mobile phone location data, which is a by-product of business operations of mobile phone companies. Such datasets often contain information about when and where different types of mobile phone activities are conducted (e.g., mobile phone calls, text messages, and location-based services). Unlike traditional travel surveys, these passively generated datasets fall short of collecting individual socioeconomic and demographic information for travel behavior and policy analysis. Moreover, individuals' location records don't explicitly reflect their travel purposes. Hence, novel approaches are needed when using such datasets for a better understanding of human mobility patterns and related transportation problems.

As suggested by Schönfelder and Axhausen (2003), an individual's home location should be treated as important anchor point when analyzing how people move around in their daily lives. Understanding how people move around their home locations could yield novel insights into people's mobility patterns. Moreover, a good understanding of people's use of space around their home locations could potentially benefit urban planning and public transport design around different neighborhoods in a city. Although passive mobile phone location data can be sparse in both space and time, it serves as a valuable data source for uncovering individual activity anchor points such as home locations, and people's use of space around these locations. Hence, this study proposes a home-based approach to studying human mobility patterns using a passive mobile phone location dataset. The main contributions of this research are as follows:

(1) With the use of a 13-day mobile phone location dataset covering more than 1 million people in the city of Shenzhen, China, we estimate each individual's "home" anchor point (which corresponds to a particular mobile phone tower) and use it as the reference point of an individual's mobility pattern. We propose a modified standard distance ($S_D{}'$) to quantify the spread of each individual's daily activities centered at the "home" anchor point. Individuals are

then grouped based on the locations of their reference points. We then derive aggregate mobility patterns for each mobile phone tower to describe how individuals who share the same "home" anchor point move around in their daily lives.

(2) Based on a multi-level hierarchical agglomerative clustering algorithm, mobile phone towers are grouped into different clusters based on their aggregate mobility patterns. Spatial distributions of the derived clusters highlight distinct human mobility patterns in different areas of the city. We then discuss the socioeconomic and demographic characteristics of the regions covered by different cluster types to gain insights of human mobility patterns in a geographical context.

The remainder of the paper is organized as follows. Section 2.2 provides a review of related work of this research. Section 2.3 introduces the mobile phone location dataset and the study area. Section 2.4 describes the methods used in this study to derive reference points of individual activity space and to analyze aggregate human mobility patterns. Section 2.5 discusses the analysis results. We summarize our findings and discuss future research directions in section 2.6.

## 2.2 Related Work

This section provides a brief review of selected studies related to the understanding of human mobility patterns, urban dynamics, individual activity anchor points using different types of tracking datasets (e.g., travel survey, GPS and mobile phone location data).

### *2.2.1 Understanding human mobility patterns*

Understanding human travel and activity patterns has always been an interest in transportation research. Before mobile technologies pervaded, travel surveys were often used in activity-based studies. Inspired by the time geography framework proposed by Hägerstrand (1970), many studies have used travel surveys to understand how people perform their daily activities that are shaped by various types of constraints. These studies covered many important subjects such as multi-purpose and individual trip chaining behavior (Hanson 1980; Kitamura 1983; Newsome et al. 1998), social roles in travel behavior (Hanson and Hanson 1980; Kwan 1999), impacts of information and communications technologies on human travel (Mokhtarian 1998; Mokhtarian 2003; Couclelis 2004), and human activity analysis in the space-time GIS framework (Kwan 2000; Miller 2005; Yu and Shaw 2008; Shaw et al. 2008; Shaw and Yu 2009; Chen et al. 2011; Yin et al. 2011). One important concept related to the understanding of human mobility patterns is individual activity space. Activity space denotes the daily environment that an individual is using for his/her activities (Golledge and Stimson 1997). There are several related concepts such as awareness space (Brown and Moore 1970), action space (Horton and Reynolds 1971), perceptual space (Relph 1976) and mental maps (Lynch 1984). In general, an individual's activity space is usually conceptualized as the locations that have been visited as well as the travel among these locations (Schönfelder and Axhausen 2003). As people's daily activities usually occur at a few locations such as home and workplaces, these locations are often considered as anchor points (Cullen and Godson 1975; Golledge and Stimson 1997) that determine the major characteristics of individual activity space. In order to describe people's daily activity patterns, considerable focus has been placed on measuring the size, geometry and structure of human activity space. For example, Dijst (1999) proposed a model with several discriminant functions to measure the size and shape of individual action space. Schönfelder and Axhausen (2003 and 2004) used three measurement approaches (confidence ellipse, kernel density estimates and minimum spanning tree) to describe the structures of people's activity spaces based on travel surveys and GPS tracking datasets. These measures can effectively describe human mobility patterns from different perspectives. However, measures like confidence ellipse often consider the arithmetic mean of individual activity locations as the center of activity space when measuring the spatial dispersion of visited locations. Schönfelder and Axhausen (2003) suggested that individuals' activity anchor points, such as home locations, should be used to substitute the arithmetic mean to "gain a behaviorally more realistic measure" (pp. 9) on human mobility patterns. This inspires us to study how people move around in their daily lives by considering their home as a key reference point.

The wide adoption of location-aware technologies such as mobile phone positioning, GPS and Wi-Fi have opened up new opportunities to study the whereabouts of people in space and time. Rhee et al. (2011) analyzed individual movements based on GPS trajectories and the study showed that human travel behavior could be described mathematically by levy-walk model. Gonzalez, Hidalgo and Barabasi (2008) and Song et al. (2010) conducted their studies based on long-term mobile phone location data and concluded that a high degree of spatiotemporal regularities existed in human movements. Kang et al. (2010) studied individual human mobility patterns among different groups divided by gender and age. Yuan, Raubal and Liu (2012) included other indicators such as mobile phone usage and transportation network densities to study how they impacted human mobility patterns. These studies yield insights into important aspects of human activity patterns. However, few of them have attempted to characterize human mobility patterns by considering important anchor points (e.g., home locations) of people's daily activities. Moreover, little has been discussed in terms of how people's mobility patterns vary geographically, which serves as important information in decision and policy making. With this in mind, our study proposes a home-based approach to studying human mobility patterns within the context of their residential locations. The proposed measure is intended to provide an intuitive way of describing the general characteristics of human activity spaces and shed light on how people move around in a geographical context.

### *2.2.2 Understanding urban dynamics using GPS and mobile phone location data*

As Batty (2009, pp. 51) pointed out in his paper, urban dynamics refers to "representations of changes in urban spatial structure through time which embody a myriad of processes at work in cities of different, but often interlocking, time scales ranging from life cycle effects in buildings and populations to movements over space and time as reflected in spatial interactions". It is apparent that human travels and activities play an important role in the manifestation of the dynamics of our cities. The Global Positioning System (GPS) has become widely adopted in studies for understanding various aspects of urban dynamics such as individual commuting patterns (Shen et al. 2013), route choice behavior (Li et al. 2005; Papinski et al. 2009) and spread of disease (Vazquez-Prokopec et al. 2009). Because of its capability to capture human movements with high spatiotemporal accuracy (Richardson et al. 2013), GPS data have been accepted as a valuable source that can help enhance our understanding of human mobility and activity patterns in urban settings (Quiroga and Bullock 1998; Bohte and Maat 2009; Bazzani et al. 2010; Shoval et al. 2011).

As an emerging technique, mobile phone positioning has become a useful way to capture human movements and activities in space and time. Much research has been done to understand various aspects of urban dynamics based on mobile phone location data. In the Real Time Rome project by MIT SENSEable City Lab, large scale anonymous cellular network data were leveraged along with instantaneous positioning of taxis and buses to portray the picture of urban mobility in real time (Calabrese and Ratti 2006; Calabrese et al. 2011). Pulselli et al. (2005), Pulselli et al. (2006) and Ratti et al. (2006) used mobile phone location data in Milan, Italy to study intensities of human activities and their changes over space and time. Several studies used mobile positioning data to study recurring patterns and daily rhythms of human movements and activities (Readers, Calabrese and Ratti 2009; Ahas et al. 2010a; Sevtsuk and Ratti 2010). In addition, mobile positioning and GPS data were used for other purposes in urban studies such as tourism analysis (Ahas et al. 2007), traffic state monitoring (Sohn and Hwang 2008), epidemiology (Tatem et al. 2009) and location recommendation (Zheng et al. 2009). In our study, we investigate human mobility patterns using a home-based approach, which can enhance our understanding of people's use of space and shed light on urban dynamics.

### *2.2.3 Identifying individual activity anchor points*

Human beings often spend a significant amount of time at specific places such as home and work locations. These key activity locations serve as important anchor points in people's everyday lives. Individual mobility patterns could be largely explained by the travel activities that occurred around these locations. Since the late 1990s, the Global Positioning System (GPS) has become a popular means of collecting tracking data for studying human travel and activity patterns. Various approaches have been applied to derive trips and important locations from individual GPS trajectories. For example, Wolf, Guensler and Bachman (2001) and Wolf et al. (2004) used GPS to collect travel data and developed several approaches to identifying trip destinations and purposes. Schuessler and Axhausen (2009) developed methods to derive individual trips and activities from GPS data. The analysis results in these studies were compared with travel diaries and demonstrated the feasibility of using GPS for understanding of individual activity patterns. Ashbrook and Starner (2003) and Zhou et al. (2007) developed algorithms to cluster GPS data points into individual activity anchor points. These studies benefited studies of human mobility patterns by uncovering their daily meaningful locations. However, GPS logs pose additional burdens on users and often suffer from problems such as signal loss (urban canyon effect) and short battery life. Mobile phones, on the other hand, become indispensable in people’s daily life and various tracking techniques deployed on the devices (GSM and Wi-Fi) make it possible to collect locations of large populations in space and time. Nurmi and Koolwaaij (2006) developed four different algorithms (graph clustering, online variant of graph clustering, spectral clustering and duration-based clustering) to derive semantics of individuals’ frequently visited locations based on information of GSM cell transitions enriched GPS data. Ahas et al. (2010b) used a passive mobile positioning dataset which covers more than half million anonymous respondents in Estonia to extract individuals’ personally meaningful locations, which “offers good potential for the monitoring of the geography and mobility of population” (p. 3). This study takes the advantage of rich information embedded in mobile phone location data to examine how people move around under the context of key activity locations.

### 2.3 Study Area and Mobile Phone Location Dataset

The study area of this research is the city of Shenzhen, China. Shenzhen is located in southern China (Figure 2.1a) and across the border from Hong Kong (Figure 2.1b). Shenzhen is China’s first Special Economic Zone (SEZ) and covers an area of 1,952 $km^2$. The city has six administrative districts: Baoan, Longgang, Nanshan, Futian, Luohu and Yantian (Figure 2.1c). As one of southern China’s financial centers, Shenzhen’s population has been growing rapidly during recent years. The urbanization process and the rapid economic growth have attracted many immigrant workers who seek job opportunities in Shenzhen. By the end of 2011, the immigrant population accounts for more than 70% of the total population in the city (see Gazette of the People’s Government of Shenzhen Municipality, 2011). The unique socioeconomic and

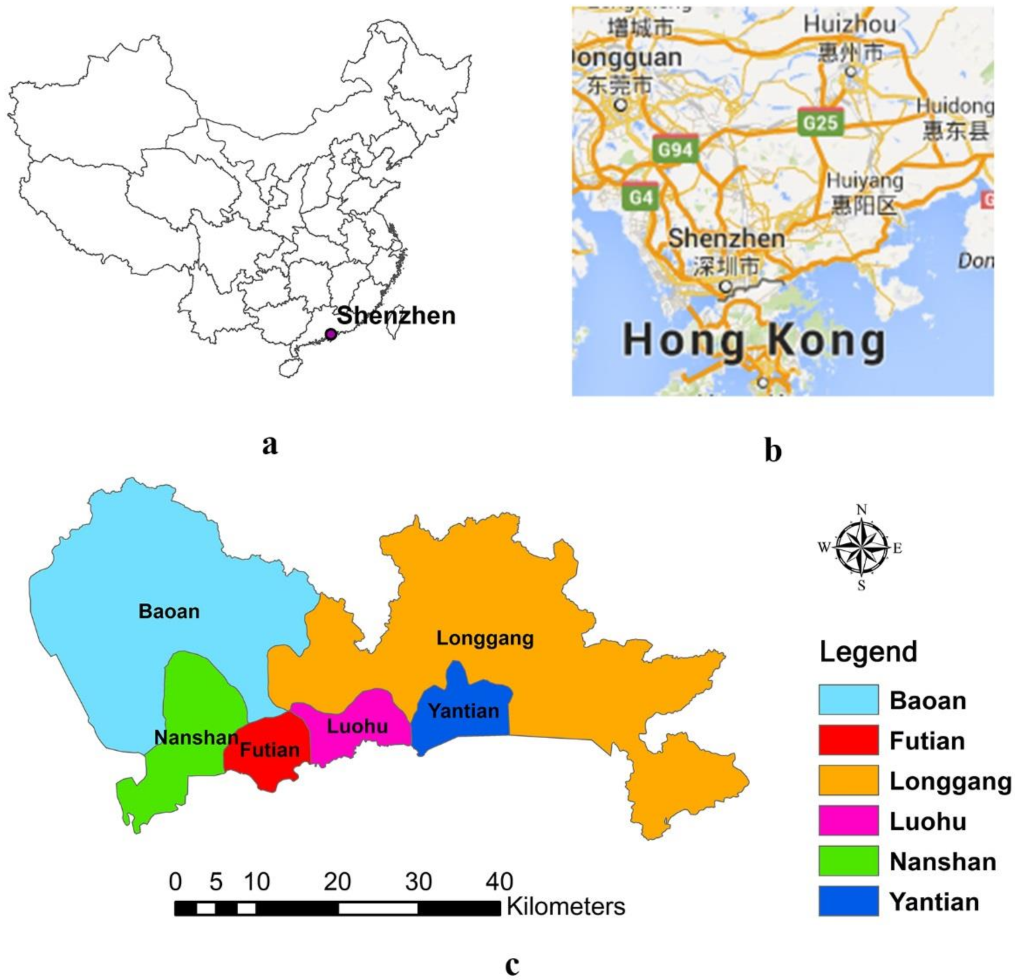



Figure 2.1 (a) Shenzhen's location in China; (b) Relative locations of Shenzhen and Hong Kong (from Google Maps); (c) Shenzhen's administrative districts.

demographic status of Shenzhen makes it an interesting area for the study of human mobility patterns.

The mobile phone location data used in this study was collected as call detail records[1] (CDRs) and each location record was generated when a mobile phone user placed or received a phone call/text message. Since such passively generated mobile phone data can be sparse in time, in this study we included 1,219,198 individuals covering a time span of 13 days with a criterion that each individual had at least 5 days with mobile phone location records. For privacy protection, this study did not obtain any personal information and each phone user in the dataset was assigned with an arbitrary user ID. In addition, all mobile phone location data were collected at the mobile phone tower level such that the specific activity locations are not revealed. As shown in Table 2.1, whenever an individual made or received a phone call or text message, a mobile phone record was generated including the user ID, date, starting time of mobile phone activity, record type (e.g., phone call vs. text message), and the coordinates (latitude/longitude) of the mobile phone tower which handled the mobile transaction. The density of mobile phone towers could vary in different parts of the study area. In this dataset, the average size of service area covered by a mobile phone tower is 0.67 $km^2$. Figure 2.2 shows the geographic distribution of all mobile phone towers included in this study. For each tower, the corresponding Thiessen Polygon is used to denote its service area.

[1] There have been a number of publications (Phithakkitnukoon et al. 2010; Song et al. 2010; Yuan et al 2012; Yuan and Raubal 2012; Becker et al. 2013) that used mobile phone location datasets for studying human mobility patterns. The mobile phone location dataset used in this study was acquired through research collaboration with Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences and the research was approved through an Institutional Review Board (IRB) process.

Table 2.1 Example of an individual user's mobile phone records during the data collection period.

| User ID | Date | Starting Time | Record Type | Longitude | Latitude |
|---|---|---|---|---|---|
| 932******** | Day 1 | 23:02:43 | Call | 113.***** | 22.***** |
| 932******** | Day 1 | 23:11:56 | Call | 113.***** | 22.***** |
| 932******** | Day 2 | 10:58:47 | Message | 113.***** | 22.***** |
| ... | … | ... | ... | 113.***** | 22.***** |
| 932******** | Day 13 | 09:22:50 | Message | 113.***** | 22.***** |

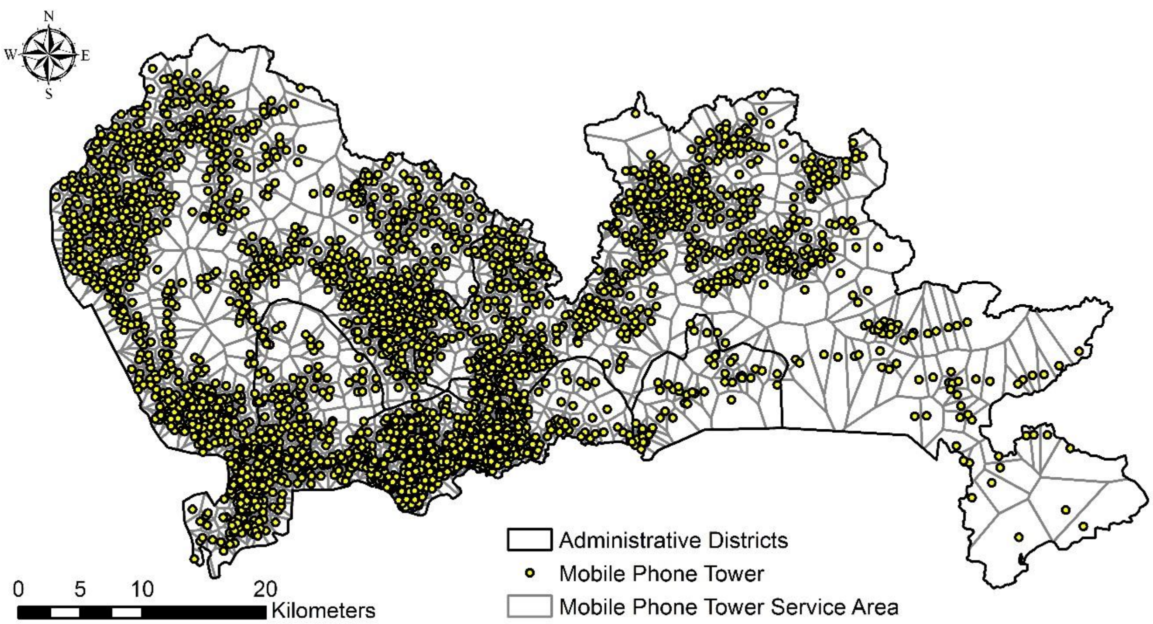



Figure 2.2 Spatial distribution of mobile phone towers in Shenzhen (2,976 in total). For each tower, the corresponding Thiessen Polygon is used to denote its service area.

## 2.4 Methodology

In this study, we propose a home-based approach to understanding aggregate human mobility patterns based on the mobile phone location dataset. First, we estimate each individual's "home" anchor point by analyzing his/her mobile phone location records. A modified standard distance measure is proposed to describe the spread of each individual's daily activities centered at the "home" anchor point during the study period. Then, each individual's estimated "home" anchor point is used as a reference point and individuals are grouped together based on the locations of their corresponding reference points. We derive aggregate mobility patterns to describe how individuals who share the same reference point move during the study period. Finally, a multi-level hierarchical agglomerative clustering algorithm is used to group the areas with similar aggregate mobility patterns into clusters. Spatial distributions of the derived clusters are then analyzed to highlight areas with distinctive aggregate mobility patterns.

### *2.4.1 Estimate individual "home" anchor point*

Places such as home and work locations serve as important anchor points of people's daily activities. This information can be acquired from travel surveys and has been used as important reference points when analyzing human activity patterns. As residential locations are not explicitly given in the anonymous mobile phone location dataset, we estimate each individual's "home" anchor point by analyzing his/her mobile phone location records. The purpose is not to pinpoint an individual's residential location, but to provide a way of measuring the spread of each individual's daily activities. Moreover, these reference points can be used as reasonable estimates of individual home locations to further investigate geographic disparities of people's travel and activity patterns.

In this study, we apply an algorithm similar to Ahas et al. (2010b) to estimate individual "home" anchor points. For each individual in the dataset, we first compute the number of days the individual has at least one record at each mobile phone tower. We then choose the top two mobile phone towers with the highest number of days as the candidate towers. The algorithm is based on an assumption that, although many people may not have many mobile phone records at the home location each day, most of them would have some mobile phone uses at the home location on a regular basis. Using the number of days with at least one record rather than the absolute number of phone records during the study period could eliminate popular locations of mobile phone usage such as transportation hubs, restaurants, and shopping malls. Finally, for each of the two candidate towers, we calculate the total number of mobile phone records during the study period and the number of mobile phone records occurred during non-work time (before 6:00 and after 18:00). The tower with higher percentage of records during non-work time is identified as the "home" anchor point.

### *2.4.2 Calculate modified standard distance*

Different measures such as standard distance (Bachi 1962) and radius of gyration (Gonzalez, Hidalgo and Barabasi 2008; Song et al. 2010) have been applied in activity-based research to describe the spread of individual activity space. These measures use the arithmetic mean (center of gravity) of each individual's activity locations as reference point. For example, the standard distance of an individual's activity space is calculated as:

$$S_D = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_c)^2}{n} + \frac{\sum_{i=1}^{n}(y_i - y_c)^2}{n}} \tag{2.1}$$

where $x_i$ and $y_i$ $(i = 1,2, \dots, n)$ denote the coordinates of an individual's sampled activity locations. $x_c$ and $y_c$ stand for the coordinates of the arithmetic mean location.

In this study, we propose a modified standard distance to measure the spread of each individual's activity space centered at his/her estimated "home" anchor point. For a particular individual in the dataset, the modified standard distance is calculated as follows:

$$S_D{}' = \sqrt{\frac{\sum_{i=1}^{n}(x_i - x_h)^2}{n} + \frac{\sum_{i=1}^{n}(y_i - y_h)^2}{n}} \tag{2.2}$$

where $n$ denotes the total number of mobile phone records for the individual. $x_i$ and $y_i$ ($i = 1,2,\ldots,n$) denote the coordinates of the $i^{th}$ mobile phone record. $x_h$ and $y_h$ are the coordinates (unit: meter) of the individual's estimated "home" anchor point under a projected coordinate system (e.g., Beijing 1954 in this study). The modified standard distance is used to reflect the spread of an individual's activities centered around the "home" anchor point. Individuals with most of their activities distributed in the vicinity of their homes tend to have smaller $S_D{}'$ than the people whose activities are more widely spread.

### *2.4.3 Derive aggregate mobility patterns for mobile phone towers*

After extracting each individual's "home" anchor point and modified standard distance, we group individuals based on the mobile phone tower that represents the location of their estimated "home" anchor point. We then analyze the aggregate mobility pattern for each mobile phone tower to distinguish the mobility patterns of people in different geographic locations of the study area.

The aggregate mobility pattern of a mobile phone tower is measured by the modified standard distances ($S_D{}'$) of the corresponding individuals. As the modified standard distance may vary significantly from person to person, using measures such as arithmetic mean may not properly reflect the characteristics of people's mobility patterns. We therefore first group the values of $S_D{}'$ into $m$ classes. Each class $D_i$ $(i = 1,2, \dots m)$ corresponds to a particular range of $S_D{}'$. The aggregate mobility pattern for a particular mobile phone tower $T$ then can be represented as a vector $V$, which consists of the percentages of individuals with their $S_D{}'$ falling in each of the $D_i$ $(i = 1,2, \dots m)$ classes:

$$V = (q_1, q_2, \dots, q_m) \tag{2.3}$$

where $q_i$ $(i = 1,2, \dots m)$ is calculated as:

$$q_i = \frac{N_i}{N} * 100\% \tag{2.4}$$

where $N$ denotes the total number of individuals with their "home" anchor point at mobile phone tower $T$. While $N_i$ $(i = 1,2, \dots m)$ denotes the number of individuals with their $S_D{}'$ falling in the class of $D_i$ $(i = 1,2, \dots m)$. Note that:

$$\sum_{i=1}^{m} N_i = N \tag{2.5}$$

$$\sum_{i=1}^{m} q_i = 1 \tag{2.6}$$

The vector $V$ is useful to distinguish the varying activity patterns of people at different locations in the study area. For example, the mobility pattern of people who live in urban centers may be different from people who live in suburban areas and could show two vectors with different characteristics.

### *2.4.4 Clustering of mobile phone towers based on aggregate mobility patterns*

Clustering multi-dimensional data can be challenging because of computational intensity and outliers. This study uses the agglomerative hierarchical clustering method with average linkage to group mobile phone towers into clusters (Han, Kamber and Pei 2011). Each cluster contains one vector initially. As shown in equation (2.7), the distance $I$ between any two vectors

$V(q_1, q_2, \ldots, q_m)$ and $V'(q_1', q_2', \ldots, q_m')$ is calculated based on the Euclidean distance measure. The similarity index between any two clusters is thus calculated as the average value of distance $I$ between vectors from the first cluster and the vectors from the second cluster. At each step, two clusters with the smallest similarity index are merged into a new cluster. The clustering process terminates when it reaches the specified number of clusters.

$$I = \sqrt{\sum_{i=1}^{m} (q_i - q_i')^2} \tag{2.7}$$

As the multidimensionality of the vectors is likely to cause the agglomerative hierarchical method to generate several small clusters with unique patterns, these outlier clusters need to be addressed during the clustering process. We adopt a multi-level clustering algorithm similar to Chen et al. (2011) to remove the outlier clusters during each iteration of the agglomerative hierarchical clustering method. As the clustering process progresses, outlier clusters are removed step by step and clusters with distinctive characteristics start to emerge.

## 2.5 Analysis Results and Discussion

This section first presents the estimation of individual "home" anchor points and their geographic distributions in the study area. Next, the modified standard distance $S_D{}'$ is calculated for each individual and the distribution of $S_D{}'$ for 1,219,198 individuals in the dataset is presented to provide an overview of the spread of people's activity space. We then group individuals based on their "home" anchor points at respective mobile phone towers to find the aggregate mobility pattern of each mobile phone tower using the method discussed in section 2.4.3. A hierarchical clustering algorithm is subsequently performed to identify clusters of mobile phone towers that share similar aggregate mobility patterns based on the method described in section 2.4.4. Spatial distributions of the derived clusters of mobile phone towers are finally mapped to highlight areas with distinctive aggregate human mobility patterns in Shenzhen.

### *2.5.1 Estimation of individual "home" anchor point*

We estimate each individual's "home" anchor point and use it as the reference point when measuring the spread of an individual's activity space. Figure 2.3 shows the density of estimated "home" anchor points of all 1,219,198 individuals in the dataset. Polygons with darker colors denote the mobile phone tower service areas with higher densities of anchor points. It can be perceived that many polygons with dark colors distribute in the southwest part, which refer to populated areas in Shenzhen. In this study, the estimated "home" anchor points are used to group individuals in space and derive aggregate human mobility patterns discussed below.

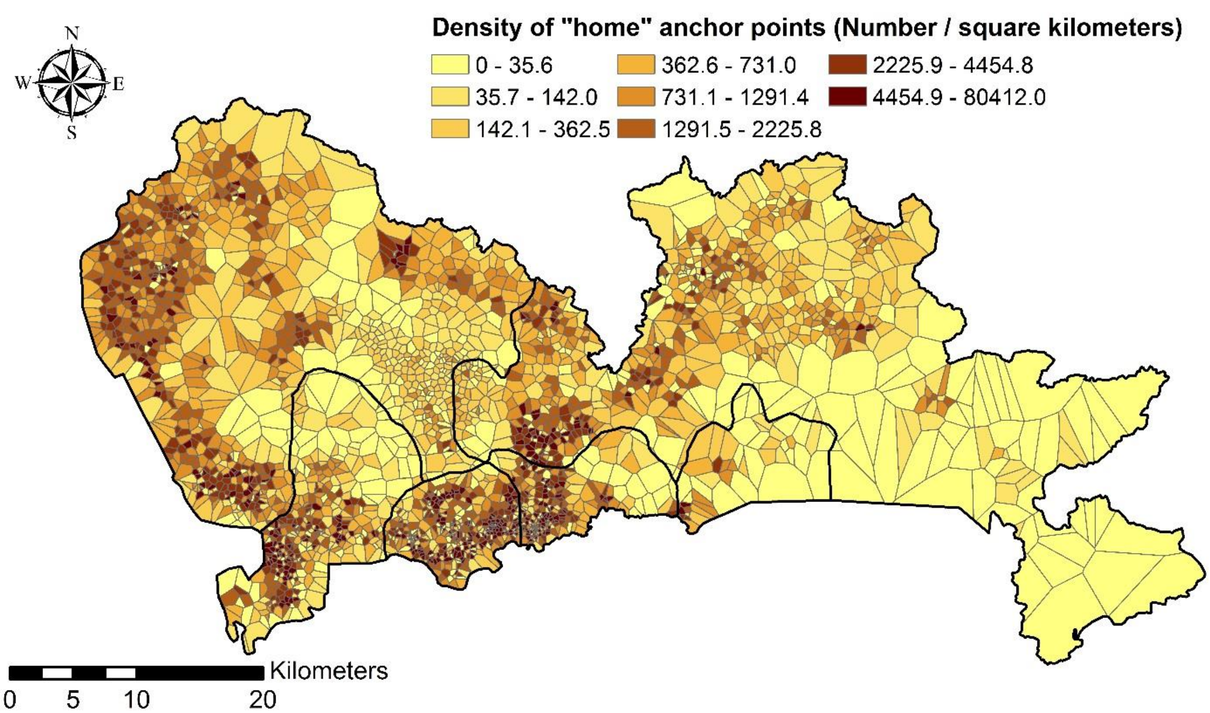



Figure 2.3 Density of estimated "home" anchor points by mobile phone tower service area.

### 2.5.2 Distribution of individual modified standard distance

With each individual's "home" anchor point, we compute the modified standard distance $S_D{}'$ to examine the spread of individual activity space centered at the "home" anchor point. Figure 2.4 shows the distribution of $S_D{}'$ for all 1,219,198 individuals in the dataset. The inserted histogram of individual $S_D{}'$ $(binwidth\ = 0.2km)$ indicates that a large percentage of the individuals has a low value of $S_D{}'$. The histogram has its peak $S_D{}'$ at $0.4 - \ 0.6\ km$, while the long tail suggests that some people have a large activity space.

The cumulative distribution in Figure 2.4 shows that 43% of the individual $S_D{}'$ are within $1km$ and 58.3% the individual $S_D{}'$ are within $2km$. On the other hand, 23.9% individual $S_D{}'$ are beyond $5km$. These statistics indicate that most people in Shenzhen have a relatively small activity space around their "home" anchor location during the 13-day study period. This encourages us to further explore the geographic landscape of people's mobility patterns.

### 2.5.3 Clustering patterns of mobile phone towers

Individuals are grouped based on their estimated "home" anchor points to derive aggregate mobility patterns by respective mobile phone towers. As mentioned in section 2.4, we divide $S_D{}'$ into several classes and apply equations (2.3) and (2.4) to measure the aggregate characteristics of human mobility patterns. According to the distribution of $S_D{}'$ shown in Figure 2.4, we divide the values of $S_D{}'$ into the following 6 classes:

(1) $\boldsymbol{D_1}$ : $0 \le S_D{}' < 1km$

(2) $\boldsymbol{D_2}$ : $1km \le S_D{}' < 2km$

(3) $\boldsymbol{D_3}$ : $2km \le S_D{}' < 3km$

(4) $\boldsymbol{D_4}$ : $3km \le S_D{}' < 4km$

(5) $\boldsymbol{D_5}$ :    $4km \leq S_D{}' < 5km$

(6) $\boldsymbol{D_6}$ :    $S_D{}' \geq 5km$

By doing so, the aggregate mobility pattern of each mobile phone tower is represented as a vector $V\ (q_1, q_2, \dots, q_6)$ recording the percentage of individuals with their $S_D{}'$ falling in each of the classes $(D_1, D_2, \dots, D_6)$. Table 2.2 gives an example of vectors for 5 randomly selected mobile phone towers. For example, the first element $(q_1)$ of mobile phone tower 1 $(T_1)$ denotes that for all individuals with their "home" anchor point at $T_1$, 27.78% of them have a $S_D{}'$ within $[0, 1km)$. The vectors can distinguish different aggregate mobility patterns associated with various mobile phone towers. For example, for $T_1$, 42.59% of the individuals have a $S_D{}'$ equal or larger than $5km$. This number drops to 17.00% when we look at $T_5$, where a large proportion of individuals exhibited low value of $S_D{}'$.

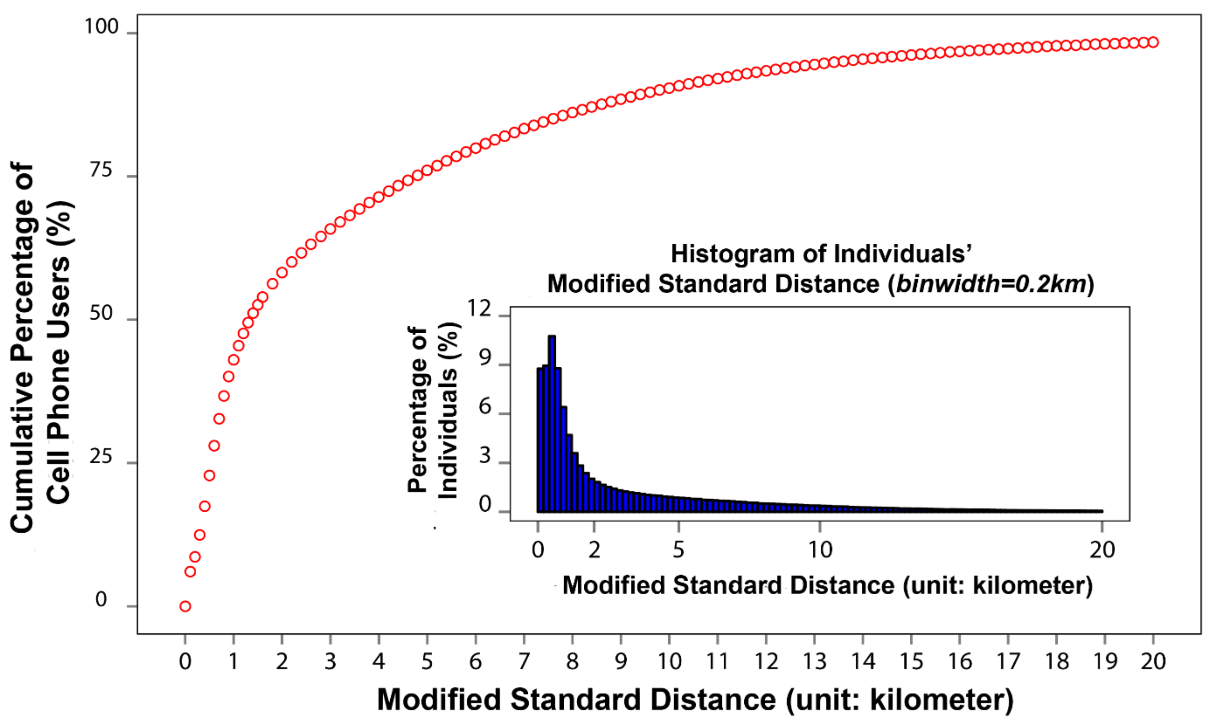



Figure 2.4 Distribution of modified standard distance $(S_D{}')$ for 1,219,198 individuals in the dataset.

Table 2.2 Vectors representing the aggregate mobility patterns of 5 randomly selected mobile phone towers.

| **Mobile Phone Tower ID** | $\boldsymbol{q_1}$ | $\boldsymbol{q_2}$ | $\boldsymbol{q_3}$ | $\boldsymbol{q_4}$ | $\boldsymbol{q_5}$ | $\boldsymbol{q_6}$ |
|---|---|---|---|---|---|---|
| $T_1$ | 27.78 % | 8.56 % | 7.34 % | 6.48 % | 7.22 % | 42.59 % |
| $T_2$ | 42.46 % | 13.51 % | 8.49 % | 6.58 % | 5.71 % | 23.22 % |
| $T_3$ | 40.31 % | 12.53 % | 9.90 % | 7.01 % | 4.11 % | 26.11 % |
| $T_4$ | 38.72 % | 11.08 % | 8.91 % | 6.21 % | 6.74 % | 28.31 % |
| $T_5$ | 49.58 % | 13.79 % | 8.79 % | 5.82 % | 4.99 % | 17.00 % |
| … | … | … | … | … | … | … |

This study performs a multi-level hierarchical agglomerative clustering algorithm on 2,976 mobile phone towers in Shenzhen that groups 2,634 mobile phone towers into 9 clusters. The remaining 342 mobile phone towers are not grouped into any cluster either because they are removed as outliers during the clustering process or there is no "home" anchor point detected at those locations. Table 2.3 shows the clustering result of these mobile phone towers. The average percentages of $V\ (q_1, q_2, \dots, q_6)$ for each cluster as well as the average percentages of $V\ (q_1, q_2, \dots, q_6)$ for all 2,976 mobile phone towers are computed for comparison purpose. The results clearly indicate distinctive characteristics of each cluster type. For example, for mobile phone towers in $C3$, a large percentage of people's $S_D{}'$ are smaller than $1km$ ($q_1$= 60.16%). Mobile phone towers in $C8$, on the other hand, have most people with $S_D{}'$ at or larger than $5km$ ($q_6$= 92.38%).

Table 2.3 Clustering results of 2,976 mobile phone towers in the dataset.

| Cluster Type | Cluster Size | Average Percentages of Vector for Each Cluster | | | | | |
|---|---|---|---|---|---|---|---|
| | | $q_1$ | $q_2$ | $q_3$ | $q_4$ | $q_5$ | $q_6$ |
| $C1$ | 617 | 27.75 % | 14.4 % | 9.57 % | 7.74 % | 6.71 % | 33.84 % |
| $C2$ | 951 | 44.3 % | 16.39 % | 7.68 % | 5.45 % | 4.29 % | 21.89 % |
| $C3$ | 496 | 60.16 % | 14.27 % | 5.51 % | 3.69 % | 2.75 % | 13.61 % |
| $C4$ | 51 | 28.65 % | 31.66 % | 10.81 % | 4.93 % | 3.41 % | 20.54 % |
| $C5$ | 28 | 15.75 % | 26.14 % | 8.49 % | 9.32 % | 6.55 % | 33.74 % |
| $C6$ | 81 | 7.05 % | 6.57 % | 4.15 % | 6.01 % | 6.77 % | 69.45 % |
| $C7$ | 165 | 15.6 % | 11.86 % | 9.08 % | 7.24 % | 7 % | 49.22 % |
| $C8$ | 227 | 0.27 % | 0.47 % | 1.06 % | 2.14 % | 3.67 % | 92.38 % |
| $C9$ | 18 | 7.11 % | 4.56 % | 16.31 % | 3.78 % | 6.93 % | 61.3 % |
| Others | 342 | - | - | - | - | - | - |
| **Overall** | **2,976** | **43%** | **15.3%** | **7.6%** | **5.6%** | **4.6%** | **23.9%** |

Figure 2.5 plots the results of these nine clusters in Table 2.3 to show their distinct patterns. Blue lines with square markers in each plot represent the average percentages of $V\ (q_1, q_2, \dots, q_6)$ for each cluster and red lines with round markers denote the average percentages of $V\ (q_1, q_2, \dots, q_6)$ for all 2,976 mobile phone towers. It is evident that, for mobile phone towers in $C1$, the percentage of people with $S_D{}' \leq 1km$ ($q_1$ =27.75%) is lower than the overall average (43%) while the percentage of people with $S_D{}' \geq 5km$ ($q_6$ =33.84%) is higher than the overall average (23.9%). $C2$ has a pattern almost identical to the overall average as it has the largest cluster size (i.e., 951). $C3$ is above the overall average of $q_1$ and below the overall average of $q_6$. $C4$ and $C5$ have dual peaks at $q_2$ and $q_6$ while having $q_1$ below the overall average. $C6$, $C7$, $C8$ and $C9$ represent the mobile phone towers with a large proportion of people having $S_D{}' \geq 5km$ and a low percentage of people having $S_D{}' < 1km$. The next section further explores the spatial distributions of these clusters to better understand people's mobility patterns in a geographical context.

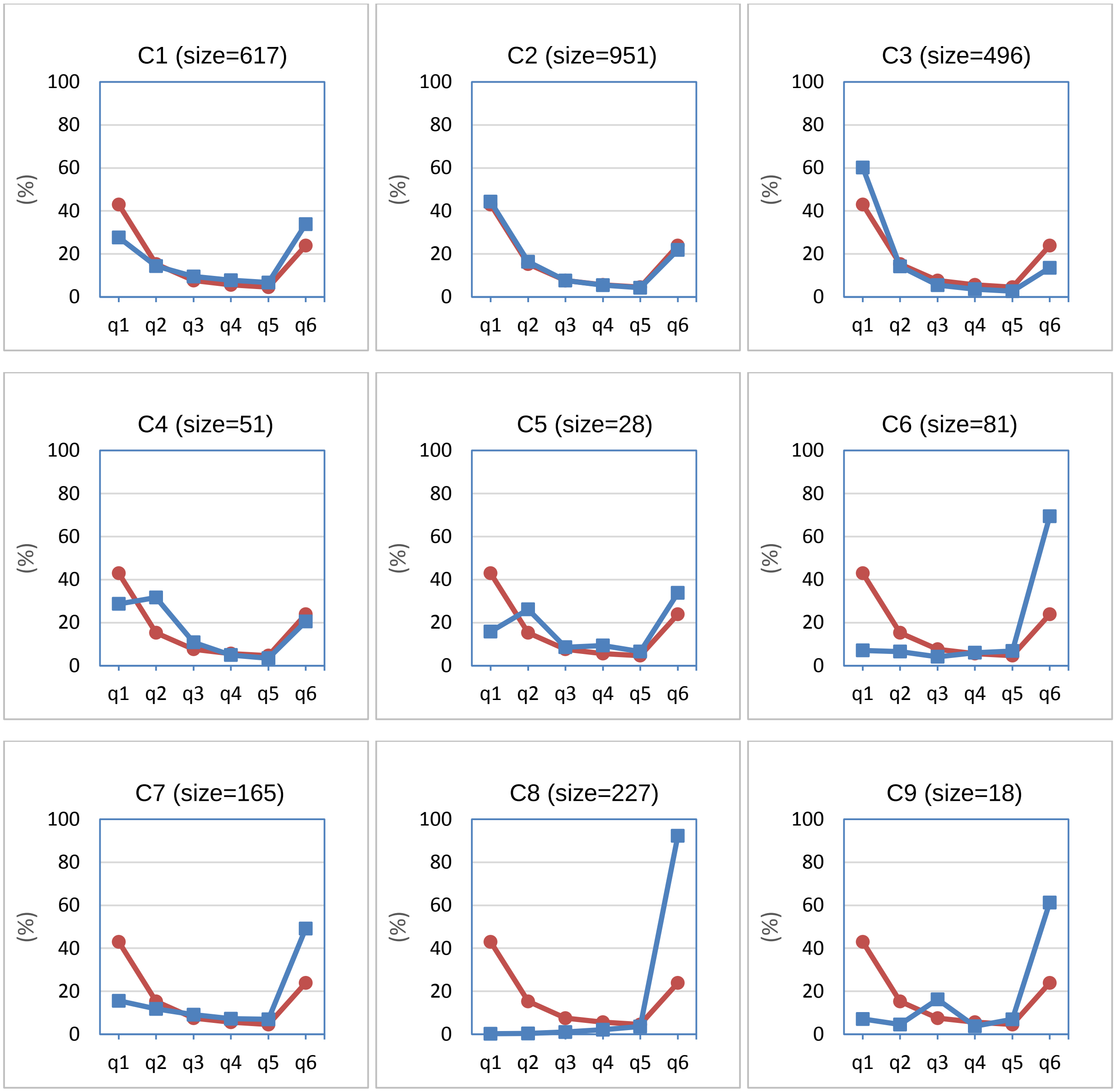



Figure 2.5 Clustering patterns of the mobile phone towers. Blue lines with square markers represent the average percentages of vector for each cluster and red lines with round markers denote the average percentage of vector for all 2,976 mobile phone towers as shown in Table 2.3.

*2.5.4 Spatial distributions of clusters with different aggregate human mobility patterns*

Spatial distributions of the above 9 clusters are mapped to further examine aggregate human mobility patterns in Shenzhen. To make the maps easier to understand, each of the maps shows the cluster(s) of similar aggregate mobility patterns. Figure 2.6 presents the spatial distribution of $C1$ with the locations of mobile phone towers represented by their service areas (i.e., Thiessen polygons in Figure 2.2). $C1$, which refers to the mobile phone towers with only 27.75% of people with $S_D' \leq 1km$ and the percentage of people with $S_D' \geq 5km$ is 33.84%, mainly covers area$s$ in the southern part of Shenzhen, with most of the mobile phone towers located in Nanshan, Futian and Luohu districts. Futian and Luohu are two major business districts where many financial and business centers are located. Nanshan, which has the highest per capita GDP among the six administrative districts in recent years, is a district with many universities and high-tech companies. In general, the economy of southern part of Shenzhen is more advanced than that of the northern part. The spatial distribution of $C1$ suggests that regional differences in economic development could be a potential factor affecting people's daily activity patterns in Shenzhen.

Compared with other clusters, $C2$ and $C3$ cover the areas where large proportion of people travelled within short distances during the study period. $C2$ has 44.3% of people with $S_D' \leq 1km$ and $C3$ has 60.16% of people with $S_D' \leq 1km$. The percentage of people with $S_D' \geq 5km$ in $C2$ and $C3$ (21.89% and 13.61% respectively) is lower than the overall average. Figure 2.7 shows the spatial distributions of $C2$ and $C3$, which cover mainly the northern part of Shenzhen in Baoan and Longgang districts. Baoan and Longgang are two industrial districts with many immigrant workers. In order to better understand the geographical context of these areas, we also display the major factories in Shenzhen in Figure 2.7. We can see that the areas covered

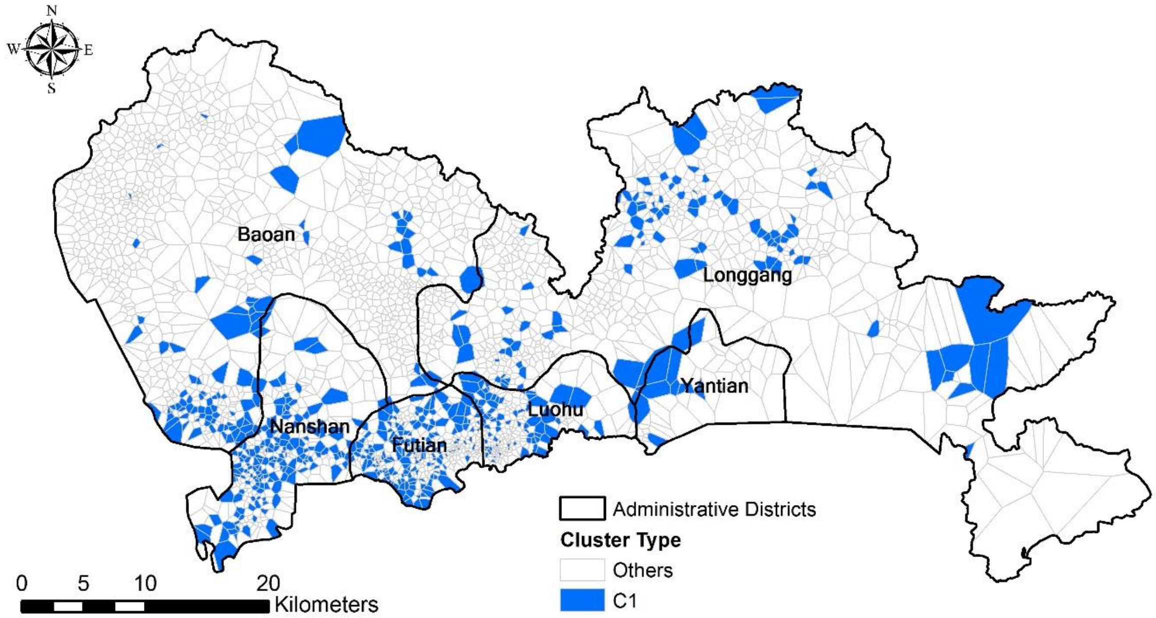



Figure 2.6 Spatial distribution of mobile phone tower service areas in $C1$.

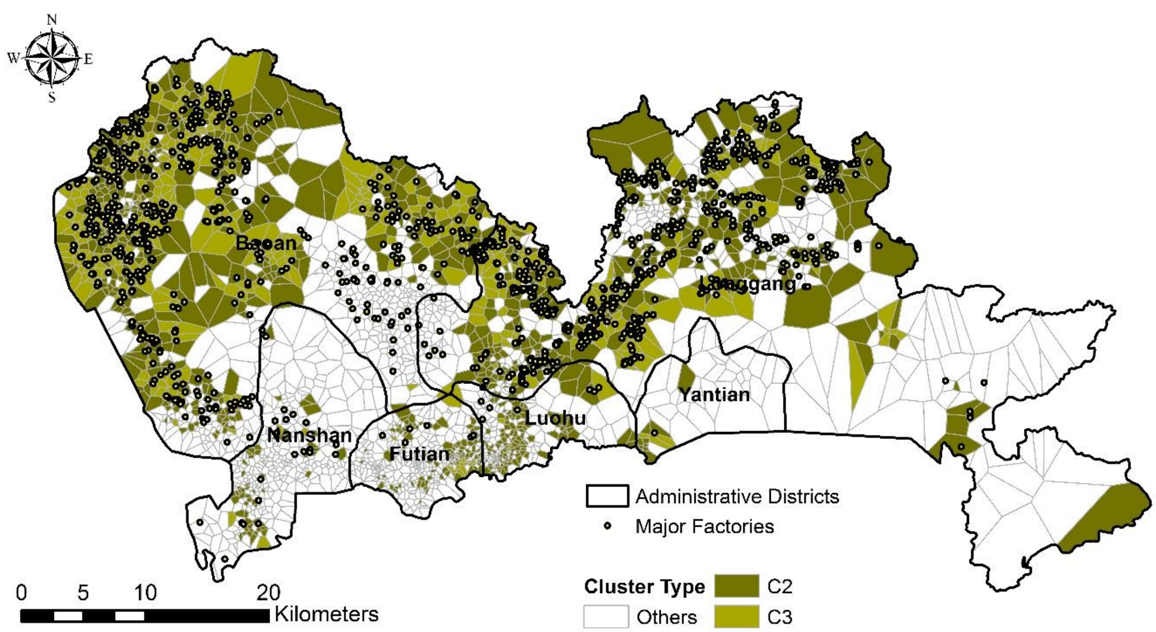



Figure 2.7 Spatial distributions of $C2$ and $C3$ and major factories in Shenzhen.

by $C2$ and $C3$ are generally co-located with the factories in Baoan and Longgang districts. Many factories in Shenzhen provide workers with dormitories adjacent to their workplace. In addition, many immigrant workers tend to rent apartments near their workplace to save commuting time and cost. We believe the low level of $S_D{}'$ in these areas could be partially explained by the daily activity patterns of many immigrant workers in these areas of Shenzhen. There are two interesting findings worth noting here: (1) some areas in $C2$ are located around the boundary between Futian and Luohu districts. It indicates that a large proportion of people in these economically developed areas also have a relatively small activity space; (2) some areas in $C2$ and $C3$ (e.g., southwestern part of Baoan and western part of Longgang) fall along the major subway lines of Shenzhen. Yet people in these areas still experience less mobility.

$C6$, $C7$, $C8$ and $C9$ cover the areas with a large percentage of people with $S_D{}' \geq 5km$. Figure 2.8 shows the spatial distributions of these four clusters. $C6$, $C7$ and $C9$ mainly locate in Nanshan, Yantian and some areas in southeast Longgang. The areas covered by $C8$ mainly locate in southeast Baoan. Table 2.3 shows that the percentage of people with $S_D{}' \geq 5km$ for $C8$ is approaching 100%. Displaying these clusters on Google Map shows that the major part of $C8$ is located in a geographically isolated area surrounded by mountains. Also, a large transportation hub (Shenzhen North Station) is located in this area, which offers transfer services to many transportation modes such as subway, taxi, bus, among others. Moreover, many inter-city highways such as Longda highway, Meiguan highway and Fulong expressway pass through this area. It should be noted that most areas covered by these four clusters are not highly populated areas in Shenzhen. One possible explanation to the observed spatial pattern is that these areas may have few job opportunities that force people to travel longer distances than people in other parts of Shenzhen. We also investigate the geographic distributions of $C4$, $C5$ as well as the outliers. Most areas covered by $C4$ are located in Baoan and Longgang districts while the areas covered by $C5$ and the outliers scatter across different districts. It appears that these clusters require additional research to identify possible causes to their spatial distribution patterns.

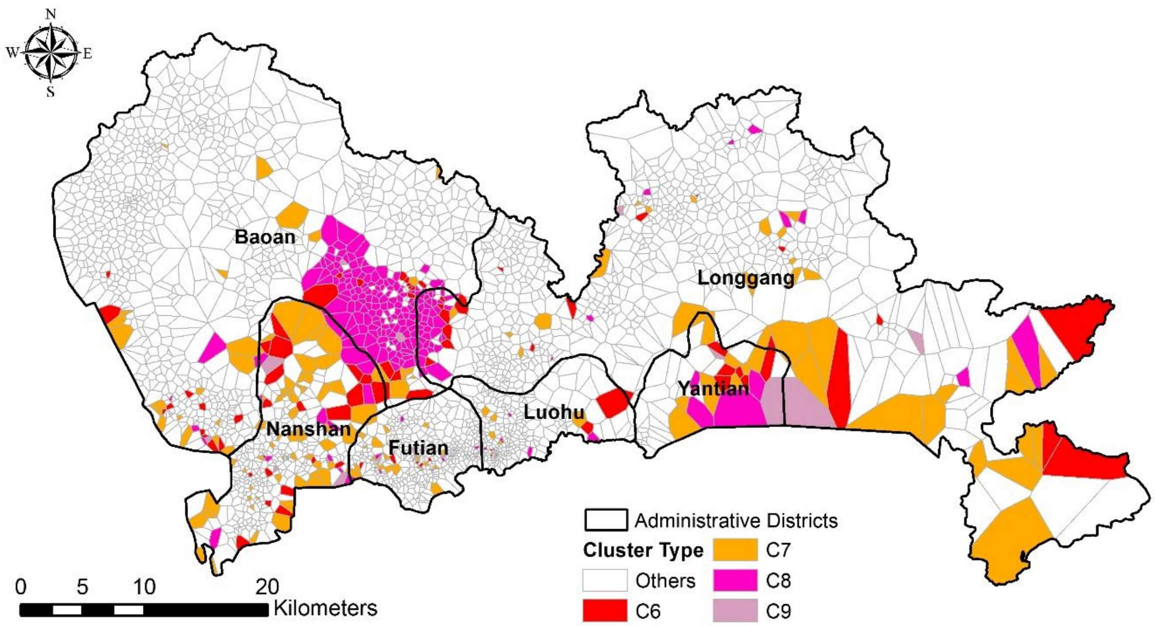



Figure 2.8 Spatial distributions of $C6$, $C7$, $C8$ and $C9$.

Major characteristics of the clusters derived from this study and their spatial distributions reveal several interesting aspects of human mobility in Shenzhen. For example, people who live around places with industrial parks and factories tend to have smaller activity space than people in other areas. One can also observe the differences of aggregate human mobility patterns between the northern and southern parts, as indicated by Figure 2.6 and Figure 2.7 respectively. The geographic differences are generally in accordance with the socio-economic divide in Shenzhen. The research findings suggest that aggregate human mobility patterns derived from mobile phone location data could serve as useful indicators for the underlying socio-demographics and land use patterns. Note that several studies have used mobile phone location data for automated land use identification (Soto and Frías-Martínez 2011; Toole et al 2012; Pei et al 2014) and the prediction of socioeconomic levels (Soto et al. 2011). The analysis results in our study yield some novel insights into people's mobility patterns in Shenzhen, and serve as an initial effort for us to understand how people use urban space and its relationship with the built environment.

## 2.6 Conclusion

The proliferation of location-aware technologies has created many passively generated datasets that track the whereabouts of people in space and time. Unlike survey or interview datasets, passively generated datasets do not explicitly report every trip and destination of the respondents during a certain time period. The snapshots of people's daily traces included in these emerging datasets encourage novel approaches for studies of human travel and activity patterns. This paper introduces a home-based approach to understanding human mobility patterns based on passive mobile phone location data. The approach considers home location as an important reference point when analyzing people's use of space. The modified standard distance is introduced to describe the spread of an individual's activity space around the home location. Individuals are then grouped in space based on their estimated home locations to derive aggregate mobility patterns at each mobile phone tower. A multi-level hierarchical clustering algorithm is performed to group areas with similar aggregate mobility patterns into clusters. The study provides a unique perspective of examining people's use of urban space with respect to his/her residential location, and complement the behavioral insights (e.g., multi-purpose, trip chaining and route choice behavior) gained from travel surveys and GPS data.

A case study using call detail records of more than 1 million mobile phones over a 13-day period collected in Shenzhen, China is carried out to test the proposed approach and methods.

The results of this case study clearly indicate the major characteristics of aggregate mobility patterns at individual mobile phone towers. The methods also identify clusters of mobile phone towers that share similar aggregate mobility patterns. Mapping the service areas of these clusters of mobile phone towers help shed light on the geographic differences of people's use of space among different parts of Shenzhen. These geographic patterns also match well with the different economic and transportation characteristics of the six districts in Shenzhen. Based on the findings, reasonable hypotheses of human travel behavior could be formulated by considering the socioeconomic and demographic characteristics of the built environment. The aggregate human mobility patterns derived at the mobile phone tower level can be further integrated with other datasets with important explanatory variables for travel behavior and policy analysis.

There are several limitations of this study. First, the passive mobile phone location data used in this study can be sparse in time since call detail records are created only when a mobile phone communication occurs. Second, this study removes those individuals who rarely use their mobile phones during the study period since it is not feasible to reliably estimate their home anchor point. The analysis results therefore do not reflect activity patterns of people who rarely use their mobile phone. Nevertheless, the mobile phone location dataset covers many people over a 13-day period in Shenzhen, China. Although some people may not be properly represented in this dataset, the analysis results offer useful insights of understanding aggregate mobility patterns in a geographic context. It is important to point out that different data sources (e.g., travel surveys, GPS, and passively generated mobile phone location data) have their respective strengths and weaknesses for studying human mobility patterns. The mobile phone location data used in this study is a valuable data source which could lead to a good understanding of people's use of space at a very large scale. However, mobile phone location data do not normally provide individual socioeconomic characteristics (e.g., age, gender, income, etc.) Studies based on mobile phone location data therefore can serve as a useful first step for more detailed and targeted follow-up studies. Moreover, while GPS data can capture locations at a finer spatial and temporal granularity but such data are not available for a large population, it would be promising to combine GPS-enabled mobile phones with passive mobile phone location data to further explore human mobility patterns.

In the future, we plan to derive several mobility indices of both spatial and temporal dimensions to describe people's mobility patterns in a more comprehensive way, and further investigate relationships of the derived indices with census (e.g., population density, gender, income, etc.) and transportation data (e.g., road networks and subway lines). This will help us better understand the intrinsic characteristics that drive people's daily travel and activities. In addition, it would be helpful to apply the same approach to similar datasets collected in other cities such that we can empirically test the robustness of the proposed approach and methods. This study demonstrates the potential of using passively generated mobile phone location data to help us gain better understanding of aggregate mobility patterns under the geographic context of a city. We believe this is only one of many possible ways of using passively generated tracking data to study human dynamics in a space-time context.

**Acknowledgement**

This study was jointly supported by National Natural Science Foundation of China (41231171), National Key Technology R&D Program of China (#2012AA12A403-4), Arts and Sciences Excellence Professorship and Alvin and Sally Beaman Professorship at the University of Tennessee, Excellent Talents Funds at Wuhan University, Shenzhen Scientific Research and Development Funding Program (ZDSY20121019111146499, JSGG20121026111056204), and Shenzhen Dedicated Funding of Strategic Emerging Industry Development Program (JCYJ20121019111128765).

## References


Ahas, R., A. Aasa, Ü. Mark, T. Pae, and A. Kull. 2007. Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management* 28 (3):898-910.

Ahas, R., A. Aasa, S. Silm, and M. Tiru. 2010a. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies* 18 (1):45-54.

Ahas, R., S. Silm, O. Järv, E. Saluveer, and M. Tiru. 2010b. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology* 17 (1):3-27.

Ashbrook, D., and T. Starner. 2003. Using GPS to learn significant locations and predict movement across multiple users. *Personal and Ubiquitous Computing* 7 (5):275-286.

Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt. 2002. Observing the rhythms of daily life: A six-week travel diary. *Transportation* 29 (2):95-124.

Bachi, R. 1962. Standard distance measures and related methods for spatial analysis. *Papers in Regional Science* 10 (1):83-132.

Batty, M. 2009. Urban modeling. *International Encyclopedia of Human Geography*, Elsevier, Oxford.

Bazzani, A., B. Giorgini, S. Rambaldi, R. Gallotti, and L. Giovannini. 2010. Statistical laws in urban mobility from microscopic GPS data in the area of Florence. *Journal of Statistical Mechanics: Theory and Experiment* 2010 (05):P05001.

Becker, R., R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. 2013. Human mobility characterization from cellular network data. *Communications of the ACM* 56 (1):74-82.

Bohte, W., and K. Maat. 2009. Deriving and validating trip purposes and travel modes for multi-day GPS-based travel surveys: A large-scale application in the Netherlands. *Transportation Research Part C: Emerging Technologies* 17 (3):285-297.

Brown, L. A., and E. G. Moore. 1970. The intra-urban migration process: a perspective. *Geografiska Annaler. Series B, Human Geography* 52 (1):1-13.

Calabrese, F., M. Colonna, P. Lovisolo, D. Parata, and C. Ratti. 2011. Real-time urban monitoring using cell phones: A case study in Rome. *Intelligent Transportation Systems, IEEE Transactions on* 12 (1):141-151.

Calabrese, F., and C. Ratti. 2006. Real time rome. *Networks and Communication Studies* 20 (3-4):247-258.

Candia, J., M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41 (22):224015.

Chen, J., S.-L. Shaw, H. Yu, F. Lu, Y. Chai, and Q. Jia. 2011. Exploratory data analysis of activity diary data: a space–time GIS approach. *Journal of Transport Geography* 19 (3):394-404.

Couclelis, H. 2004. Pizza over the Internet: e-commerce, the fragmentation of activity and the tyranny of the region. *Entrepreneurship & Regional Development* 16 (1):41-54.

Cullen, I., and V. Godson. 1975. Urban networks: the structure of activity patterns. *Progress in planning* 4:1-96.

Dijst, M. 1999. Two-earner families and their action spaces: A case study of two Dutch communities. *GeoJournal* 48 (3):195-206.

Gazette of the People’s Government of Shenzhen Municipality. Issue No. 17, Serial No. 741, April 22, 2011. <http://www.sz.gov.cn/zfgb/2012_1/gb785/201204/t20120423_1844697.htm>.

Golledge, R., and R. Stimson. 1997. Spatial behaviour. *Guilford, London*.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196):779-782.

Hägerstraand, T. 1970. What about people in regional science? *Papers in Regional Science* 24 (1):7-24.

Han, J., M. Kamber, and J. Pei. 2011. *Data mining: concepts and techniques: concepts and techniques*. Elsevier.

Hanson, S. 1980. The importance of the multi-purpose journey to work in urban travel behavior. *Transportation* 9 (3):229-248.

Hanson, S., and P. Hanson. 1980. Gender and urban activity patterns in Uppsala, Sweden. *Geographical Review*:291-299.

Horton, F. E., and D. R. Reynolds. 1971. Effects of urban spatial structure on individual behavior. *Economic Geography*:36-48.

Kang, C., S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, and X. Ma. 2010. Analyzing and geo-visualizing individual human mobility patterns using mobile call records. In Proceedings of 18th International Conference on Geoinformatics (pp. 1-7). IEEE.

Kitamura, R. 1984. Incorporating trip chaining into analysis of destination choice. *Transportation Research Part B: Methodological* 18 (1):67-81.

Kwan, M. P. 1999. Gender, the Home-Work Link, and Space-Time Patterns of Nonemployment Activities. *Economic Geography* 75 (4):370-394.

Kwan, M.-P. 2000. Interactive geovisualization of activity-travel patterns using three-dimensional geographical information systems: a methodological exploration with a large data set. *Transportation Research Part C: Emerging Technologies* 8 (1):185-203.

Levinson, D., and A. Kumar. 1995. Activity, travel, and the allocation of time. *Journal of the American Planning Association* 61 (4):458-470.

Li, H., R. Guensler, and J. Ogle. 2005. Analysis of morning commute route choice patterns using global positioning system-based vehicle activity data. *Transportation Research Record: Journal of the Transportation Research Board* 1926 (1):162-170.

Lynch, K. 1984. *Good city form*: MIT press, Cambridge.

Miller, H. J. 2005. A measurement theory for time geography. *Geographical Analysis* 37 (1):17-45.

Mokhtarian, P. L. 1998. A synthetic approach to estimating the impacts of telecommuting on travel. *Urban Studies* 35 (2):215-241.

Mokhtarian, P. L. 2002. Telecommunications and travel: The case for complementarity. *Journal of Industrial Ecology* 6 (2):43-57.

Newsome, T. H., W. A. Walcott, and P. D. Smith. 1998. Urban activity spaces: Illustrations and application of a conceptual model for integrating the time and space dimensions. *Transportation* 25 (4):357-377.

Nurmi, P., and J. Koolwaaij. 2006. Identifying meaningful locations. *3rd Annual International Conference on Mobile and Ubiquitous Systems: Networking & Services, (*pp. 1-8). San Jose, California: IEEE.

Papinski, D., D. M. Scott, and S. T. Doherty. 2009. Exploring the route choice decision-making process: A comparison of planned and observed routes obtained using person-based GPS. *Transportation research part F: traffic psychology and behaviour* 12 (4):347-358.

Pei, T., S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou. 2014. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science* 28(9): 1997-2007.

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*, 14-25: Springer.

Pulselli, R., F. Pulselli, C. Ratti, and E. Tizzi. 2005. Dissipative structures for understanding cities: Resource flows and mobility patterns. In Proceedings of the First International Conference on Built Environment Complexity. pp. 271-279.

Pulselli, R., C. Ratti, and E. Tiezzi. 2006. City out of chaos: social patterns and organization in urban systems. *International Journal of Ecodynamics* 1(2):125-134.

Quiroga, C. A., and D. Bullock. 1998. Travel time studies with global positioning and geographic information systems: an integrated methodology. *Transportation Research Part C: Emerging Technologies* 6 (1):101-127.

Ratti, C., S. Williams, D. Frenchman, and R. Pulselli. 2006. Mobile landscapes: using location data from cell phones for urban analysis. *Environment and Planning B: Planning and Design* 33 (5):727.

Reades, J., F. Calabrese, and C. Ratti. 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design* 36 (5):824-836.

Relph, E. 1976. *Place and placelessness*: Pion London.

Rhee, I., M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. 2011. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)* 19 (3):630-643.

Richardson, D. B., N. D. Volkow, M.-P. Kwan, R. M. Kaplan, M. F. Goodchild, and R. T. Croyle. 2013. Spatial turn in health research. *Science* 339 (6126):1390.

Schönfelder, S., and K. W. Axhausen. 2003. Activity spaces: measures of social exclusion? *Transport Policy* 10 (4):273-286.

Schönfelder, S., and K. Axhausen. 2004. Structure and innovation of human activity spaces. *Arbeitsberichte Verkehrs-und Raumplanung* 258.

Schuessler, N., and K. W. Axhausen. 2009. Processing raw data from global positioning systems without additional information. *Transportation Research Record: Journal of the Transportation Research Board* 2105 (1):28-36.

Sevtsuk, A., and C. Ratti. 2010. Does urban mobility have a daily routine? Learning from the aggregate data of mobile networks. *Journal of Urban Technology* 17 (1):41-60.

Shaw, S. L., H. Yu, and L. S. Bombom. 2008. A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12 (4):425-441.

Shaw, S.-L., and H. Yu. 2009. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical–virtual space. *Journal of Transport Geography* 17 (2):141-149.

Shen, Y., M.-P. Kwan, and Y. Chai. 2013. Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. *Journal of Transport Geography* 32:1-11.

Shoval, N., H.-W. Wahl, G. Auslander, M. Isaacson, F. Oswald, T. Edry, R. Landau, and J. Heinik. 2011. Use of the global positioning system to measure the out-of-home mobility of older adults with differing cognitive functioning. *Ageing and Society* 31 (05):849-869.

Sohn, K., and K. Hwang. 2008. Space-based passing time estimation on a freeway using cell phones as traffic probes. *Intelligent Transportation Systems, IEEE Transactions on* 9 (3):559-568.

Song, C., Z. Qu, N. Blumm, and A.-L. Barabási. 2010. Limits of predictability in human mobility. *Science* 327 (5968):1018-1021.

Soto, V., and E. Frías-Martínez. 2011. Automated land use identification using cell-phone records. Paper read at Proceedings of the 3rd ACM international workshop on MobiArch. pp. 17-22.

Soto, V., V. Frias-Martinez, J. Virseda, and E. Frias-Martinez. 2011. Prediction of socioeconomic levels using cell phone records. In *User Modeling, Adaption and Personalization*, 377-388: Springer.

Tatem, A. J., Y. Qiu, D. L. Smith, O. Sabot, A. S. Ali, and B. Moonen. 2009. The use of mobile phone data for the estimation of the travel patterns and imported Plasmodium falciparum rates among Zanzibar residents. *Malar J* 8:287.

Toole, J. L., M. Ulm, M. C. González, and D. Bauer. 2012. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, pp. 1-8.

Vazquez-Prokopec, G. M., S. T. Stoddard, V. Paz-Soldan, A. C. Morrison, J. P. Elder, T. J. Kochel, T. W. Scott, and U. Kitron. 2009. Usefulness of commercially available GPS data-loggers for tracking human movement and exposure to dengue virus. *International Journal of Health Geographics* 8 (1):68.

Wolf, J., R. Guensler, and W. Bachman. 2001. Elimination of the travel diary: Experiment to derive trip purpose from global positioning system travel data. *Transportation Research Record: Journal of the Transportation Research Board* 1768 (1):125-134.

Wolf, J., S. SchöUnfelder, U. Samaga, M. Oliveira, and K. Axhausen. 2004. Eighty weeks of global positioning system traces: approaches to enriching trip information. *Transportation Research Record: Journal of the Transportation Research Board* 1870 (1):46-54.

Yin, L., S.-L. Shaw, and H. Yu. 2011. Potential effects of ICT on face-to-face meeting opportunities: a GIS-based time-geographic approach. *Journal of Transport Geography* 19 (3):422-433.

Yu, H., and S. L. Shaw. 2008. Exploring potential human activities in physical and virtual spaces: a spatio-temporal GIS approach. *International Journal of Geographical Information Science* 22 (4):409-430.

Yuan, Y., and M. Raubal. 2012. Extracting dynamic urban mobility patterns from mobile phone data. In *Geographic Information Science*, 354-367: Springer.

Yuan, Y., M. Raubal, and Y. Liu. 2012. Correlating mobile phone usage and travel behavior–A case study of Harbin, China. *Computers, Environment and Urban Systems* 36 (2):118-130.

Zheng, Y., L. Zhang, X. Xie, and W.-Y. Ma. 2009. Mining interesting locations and travel sequences from GPS trajectories. In Proceedings of the 18th international conference on World wide web pp. 791-800.

Zhou, C., D. Frankowski, P. Ludford, S. Shekhar, and L. Terveen. 2007. Discovering personally meaningful places: An interactive clustering approach. *ACM Transactions on Information Systems (TOIS)* 25 (3):12.

# CHAPTER 3

# ANOTHER TALE OF TWO CITIES – UNDERSTANDING HUMAN ACTIVITY SPACE USING ACTIVELY TRACKED CELLPHONE LOCATION DATA

This chapter is a slightly modified version of a manuscript submitted to the *Annals of the Association of American Geographers*. The manuscript has been accepted and will be published in the 2016 special issue on "Geographies of Mobilities". The use of "we" in this chapter refers to co-authors, Dr. Shih-Lung Shaw, Ziliang Zhao, Dr. Ling Yin, Dr. Feng Lu, Dr. Jie Chen, Dr. Zhixiang Fang, Dr. Qingquan Li, and myself. As the first author, I processed the data, performed the analysis and wrote the manuscript.

**Abstract:** Activity space is an important concept in geography. Recent advancements of location-aware technologies have generated many useful spatiotemporal datasets for studying human activity space over a large population. In this paper, we use two active tracking cellphone location datasets that cover a weekday to characterize people's use of space in Shanghai and Shenzhen, China. We introduce three mobility indicators (daily activity range, number of activity anchor points, and frequency of movements) to represent the major determinants of individual activity space. By applying association rules in data mining theory, we analyze how these indicators combine with each other in each individual's activity space to gain insights of mobility patterns in these two cities. We further examine spatiotemporal variations of aggregate mobility patterns in these two cities. Our analysis results reveal some distinctive characteristics of human activity space in these two cities: (1) A high percentage of people in Shenzhen has a relatively short daily activity range, while people in Shanghai exhibit a variety of daily activity ranges; (2) People with more than one activity anchor points tend to travel further, but less frequently, in Shanghai than in Shenzhen; (3) Shenzhen shows a significant "north-south" contrast of activity space which reflects the urban structure of Shenzhen; (4) Travel distance in both cities is lower around noon than regular work hours, and a large percentage of the movements around noon are associated with individual home locations. This study indicates the benefits of analyzing active tracking cellphone location data to gain insights of human activity space in different cities.

**Keywords:** human activity space, association rules, spatiotemporal patterns, active cellphone location data

## 3.1 Introduction

Human activities and movements generate the pulses of our cities. Studying human activity space could yield important insights into many socioeconomic phenomena, and facilitate our understanding of human behavior and its relationships with the built environment. Activity space is an important concept in geography that describes the spatial extent, frequent locations, and movements of people's daily activities (Golledge and Stimson 1997; Schönfelder and Axhausen 2003). In the past several decades, studies of human activity space were mainly based on travel surveys and GPS data (Hanson 1980; Dijst 1999; Kwan 1999; Axhausen et al. 2002; Shoval and Isaacson 2007; Zheng et al. 2008; Shaw, Yu and Bombom 2008; Chen et al. 2011; Shen, Kwan and Chai 2013). Recent advancements of location-aware technologies have made it possible to collect large-scale individual tracking datasets for studying the whereabouts of people over space and time. These newly emerged data sources, like social media and cellphone location data, provide us with opportunities to investigate human activity space over a large population. Although various methods have been suggested to measure people's use of space (Candia et al. 2008; Isaacman et al. 2010; Song 2010; Cheng et al. 2011; Cho, Myers and Leskovec 2011; Becker et al. 2013; Silm and Ahas 2014), several research challenges remain. For example, many previous studies examined important determinants of human activity space separately. It remains unclear how different determinants of an individual activity space are related to each other. Although some studies have used methods such as clustering to identify mobility patterns based on multiple characteristics of individual activity space, it can be difficult to interpret the major characteristics of each population group. In this study, we develop some intuitive individual mobility indicators (IMIs) to represent individual activity space from three critical perspectives (i.e., spatial extent, frequent locations, and movements). We then introduce several approaches to uncover the interrelationships of these mobility indicators, and compare activity space patterns among different cities or population groups.

We use two large-scale active tracking cellphone location datasets collected in two major Chinese cities, Shanghai and Shenzhen, on a workday to investigate and compare human activity space patterns between these two cities as an example to illustrate the usefulness of our proposed individual mobility indicators. Different from call detail records (CDRs) that are passively collected when people engage in communication activities such as phone calls and text messages (Song 2010; Becker et al. 2013; Xu et al. 2015), active tracking cellphone location data provide locations of each cellphone at a regular time interval by detecting where a cellphone is located. Since many people make infrequent use of their cellphone in a day and cellphone usage tends to have a natural biased spatiotemporal pattern (e.g., more cellphone communications after work than before work in a day), active tracking cellphone location data generally offers a better spatiotemporal coverage of individual activity space than CDR data. The main objective of this paper is to develop a method that can measure the major characteristics of individual activity space based on active tracking cellphone location data such that we can effectively compare aggregate activity space patterns among different cities. To achieve the objective, we develop three individual mobility indicators, which are daily activity range, number of activity anchor points and frequency of movement, to answer critical questions of individual activity space (i.e., how far, how many, and how frequent). We then apply association rules in data mining (Han, Kamber and Pei 2011) to examine how the three indicators are related to each other among the activity spaces of different individuals. We further investigate spatial and temporal variations of major characteristics of aggregate human activity patterns between Shanghai and Shenzhen.

## 3.2 Literature Review

Activity space and its related concepts (Lynch 1960; Brown and Moore 1970; Horton and Reynolds 1971; Lenntorp 1976; Golledge and Stimson 1997) have been widely used in geography to examine people's use of space. Various approaches including, but not limited to, standard deviational ellipse (Yuill 1971), confidence ellipse (Schönfelder and Axhausen 2003) and daily potential path area (Kwan 1999) have been proposed to measure individual space usage from perspectives of spatial extent, frequent locations, and movements. Over the past several decades, many studies have applied these approaches to study human activity space and its relationships with socio-demographic characteristics (Hanson and Hanson, 1981; Newsome, Walcott and Smith 1998; Dijst 1999; Kwan 1999; Axhausen et al. 2002; Buliung and Kanaroglou 2006). Most of these studies involved activity/travel surveys that can be expensive to collect and often limited in sample size. As we move into the big data era, many new data sources have emerged. For example, there have been several studies which used active tracking mobile phone location data to solve problems related to mobility prediction (Gao, Tang and Liu 2012), recognition of place categories (Zhu et al. 2012), and estimation of demographic attributes (Brdar, Culibrk and Crnojevic 2012). Such datasets provide new opportunities to the understanding of people's use of space in their daily lives. However, large data volume presents new challenges to the study of human activity space. In recent years, research has been conducted to study human activity space using cellphone location data. Measures such as radius of gyration (Gonzalez, Hidalgo and Barabási 2008; Song, Blumn and Barabási 2010), activity anchor points (Phithakkitnukoon et al 2010; Cho, Myers and Leskovec 2011) and daily activity range (Becker et al 2013) have been used to reflect major characteristics of individual activity space. However, most of the studies analyzed these characteristics separately, which could lead to a partial view of individual activity space.

Although clustering methods have been applied to address some research issues, such as identifying individuals with similar location sequence (Li et al. 2008), commuting flexibility (Shen, Kwan and Chai 2013) and spatiotemporal activity patterns (Chen et al. 2011), it sometimes can be difficult to interpret the major characteristics of each population group derived from the clustering algorithms. Moreover, these clustering methods (e.g., hierarchical clustering) are computation-intensive and often perform inefficiently over very large datasets. This study attempts to develop some easy-to-compute, yet effective approaches to gain insights of activity space patterns. We build three mobility indicators to represent the most important determinants of individual activity space. By combining activity space theory and association rules in data mining, this study aims at providing a multi-dimensional view of individual activity space, and facilitating the comparison of human activity spaces across different cities.

## 3.3 Study Area and Datasets

Shanghai and Shenzhen are two major cities in China with their Gross Domestic Product (GDP) ranked the 1st and the 4th respectively among all Chinese cities (National Bureau of Statistics of China, 2012). Shanghai is a century-old metropolis, with a population of 24 million as of 2013. It has 18 administrative districts and covers an area of 6,340 $km^2$ (Figure 3.1a and 3.1b). Shenzhen, which is located in southern China adjacent to Hong Kong, has 6 administrative districts covering 1,952 $km^2$ and an estimated population of 15 million (Shenzhen Daily 2012) as of 2012 (Figure 3.1c). Shenzhen was a small fishing village when it was chosen as China's first Special Economic Zone (SEZ) in 1979. Fast economic growth and urbanization have transformed Shenzhen into a major migrant city. As of 2011, the migrant population accounted for more than 70 percent of the total population in Shenzhen (Gazette of the People's Government of Shenzhen Municipality, 2011). According to recent travel surveys (Lu and Gu 2011; Transport Commission of Shenzhen Municipality 2011), non-motorized trips accounted for a large percentage of total trips in Shanghai (walking: 26.2% and bicycle/moped: 28.7%) and in Shenzhen (walking: 50.0% and bicycle/moped: 6.2%). Comparing people's daily activity space in these two cities can help us better understand their urban dynamics that could be useful for urban design, transportation planning, business studies, among other applications.

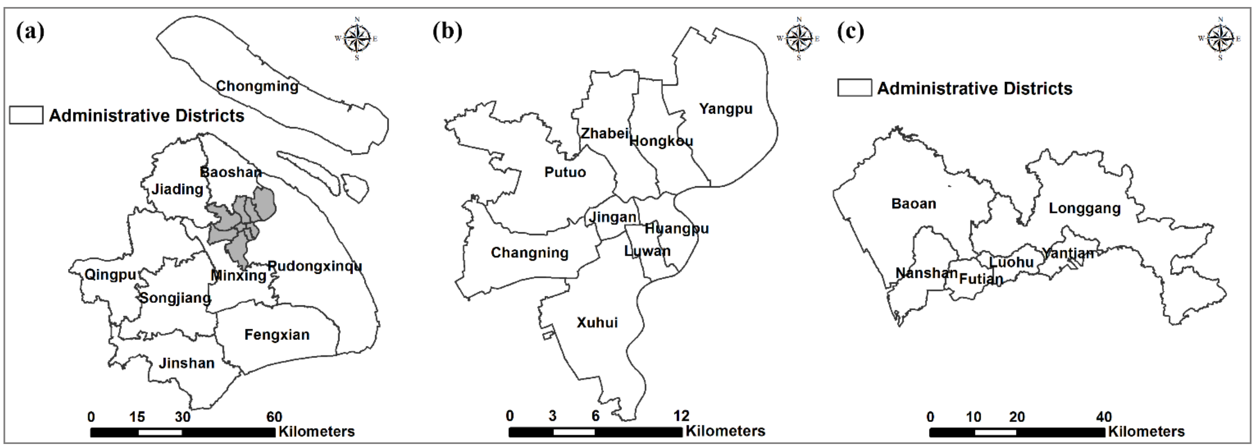



Figure 3.1 Study areas: (a) administrative districts of Shanghai; (b) inset map of the central part of Shanghai; (c) administrative districts of Shenzhen.

This article uses two active tracking cellphone datasets[1] collected on a weekday in Shenzhen (03/23/2012) and Shanghai (09/03/2012), respectively. The Shenzhen dataset covers 5.8 million cellphones, with their locations reported approximately once every hour as (x, y)

[1] The mobile phone location dataset used in this study was acquired through research collaboration with the Shenzhen Institute of Advanced Technology, Chinese Academy of Sciences and the research was approved by Institutional Review Board (IRB).

coordinates of the cellphone tower to which a cellphone is assigned. This dataset does not include data for the 23:00–24:00 time window; each cellphone therefore has 23 observations in a day. The Shanghai dataset consists of 0.69 million cellphones. To be comparable, we removed records of the 23:00–24:00 time window in Shanghai's dataset. Table 3.1 shows an example of the data format of the two cellphone datasets. The average nearest distance among the cellphone towers in Shanghai is $0.21km$, as compared to $0.19km$ in Shenzhen.

Table 3.1 Example of an individual's cellphone records in both datasets.

| **User ID** | **Record ID** | **Time Window in which location was reported ($t$)** | **Longitude of cellphone tower ($x$)** | **Latitude of cellphone tower ($y$)** |
|---|---|---|---|---|
| 932***** | 1 | [00:00 – 01:00) | 113.***** | 22.***** |
| 932***** | 2 | [01:00 – 02:00) | 113.***** | 22.***** |
| 932***** | 3 | [02:00 – 03:00) | 113.***** | 22.***** |
| **...** | **...** | **...** | 113.***** | 22.***** |
| 932***** | 23 | [22:00 – 23:00) | 113.***** | 22.***** |

## 3.4 Methodology

This section first introduces three individual mobility indicators (IMIs), followed by estimation of each individual's home location that will be used as a reference point when we analyze individual activity space. We then describe how association rules are used to summarize and compare people's activity space in Shanghai and Shenzhen.

### *3.4.1 Individual Mobility Indicators*

As shown in Table 3.1, an individual's cellphone trajectory $T$ can be represented as:

$$T = \{ P_1(x_1, y_1, t_1), P_2(x_2, y_2, t_2), \ldots, P_i(x_i, y_i, t_i) \} \quad (3.1)$$

where $P_i$ denotes the $i^{th}$ $(i = 1,2, \ldots ,23)$ cellphone location record; $x_i$ and $y_i$ denote the coordinates (unit: meter) of a cellphone tower under a projected coordinate system (e.g., Beijing 1954 in this study), and $t_i$ represents an one-hour time window in which each location was recorded. We develop three individual mobility indicators (IMIs), which are the number of activity anchor points, daily activity range, and frequency of movement, to capture the major characteristics of an individual activity space represented by $T$.

Measures such as standard deviational ellipse (Yuill 1971) and radius of gyration (Gonzalez, Hidalgo and Barabási 2008) have been used in previous studies to represent the spatial dispersion of an individual's daily activities. In our study, we introduce *daily activity range*, which is defined as the maximum distance between all pairs of cellphone towers in $T$, to describe the spatial extent of an individual's activity space[2].

Activity anchor point has been frequently used in the literature (e.g., Dijst 1999; Schönfelder and Axhausen 2003; Ahas et al. 2010) to denote a person's major activity locations such as home, workplace, favorite restaurants, etc. However, the meaning of activity anchor point could vary due to the context of each study. In this paper, we define an *activity anchor point* as a set of cellphone towers that are geographically concentrated and where an individual spent a certain amount of time. One challenge of using cellphone location data to determine an individual's activity anchor points and movements among the anchor points is that an individual's cellphone location could switch among adjacent cellphone towers due to either cellphone load balancing (Csáji et al. 2013) or cellphone signal strength variation (Isaacman et al. 2012). Hence, in order to

[2] Radius of gyration (Gonzalez, Hidalgo and Barabási 2008) is another frequently used measure that describes the spatial dispersion of an individual's activity space. In this study, we calculate both daily activity range and radius of gyration, and we find that they are highly correlated with each other (Pearson Coefficient = 0.96). We thus use daily activity range in this study to represent the spatial extent of an individual's activity space due to its intuitive meaning.

derive *activity anchor points* for $T$, we first extract all cellphone towers traversed by $T$, and calculate the frequency (i.e., number of time windows) each cellphone tower was visited. We then select the most visited cellphone tower, and group all the cellphone towers that are located within $0.5km$ of the selected tower into a cluster. We then select the next most visited cellphone tower and perform the same grouping process. The process is repeated until all the cellphone towers in $T$ are processed. Finally, we calculate the number of cellphone location records (i.e., observations) assigned to each cluster. In this study, any cluster with two or more cellphone location records is identified as an *activity anchor point*. Figure 3.2 gives an example of an individual's trajectory in a 3-dimensional space-time system proposed by Hägerstrand (1970). This individual's cellphone tower locations are grouped into four *clusters*, with three of them (clusters A, B and C) being identified as *activity anchor points*.

Note that we choose a constant distance threshold of $0.5km$ to derive individual Activity Anchor Points for the two cellphone datasets, and the reasons are as follows. First, although we are aware that cellphone tower densities could vary within a city, choosing a constant threshold enables us to consistently evaluate the space usage of individuals in a city. Second, as Shanghai and Shenzhen share a similar average nearest distance among cellphone towers ($0.21km$ and $0.19km$, respectively), choosing $0.5km$ can not only address the problem of signal switches among nearby cellphone towers but also facilitate the comparison of human activity space between these two cities.

Movement is another important characteristic of human activity space. When deriving *frequency of movement* in $T$, we only consider the movements occurred between clusters (green lines in Figure 3.2) since it is difficult to determine whether the movements within clusters (i.e., red lines in Figure 3.2) represent an individual's actual movements, or they are simply caused by load balancing or signal switches. By choosing a threshold of 0.5km, we minimize the impact of load balancing and signal switches while maintaining all major movements in an individual's trajectory. Here, the *frequency of movement* is defined as the number of inter-cluster movements in $T$. This indicator measures how active an individual travels between different activity anchor locations in a day. For example, the individual in Figure 3.2 has a *frequency of movement* of 5 (i.e., the number of green segments in Figure 3.2).

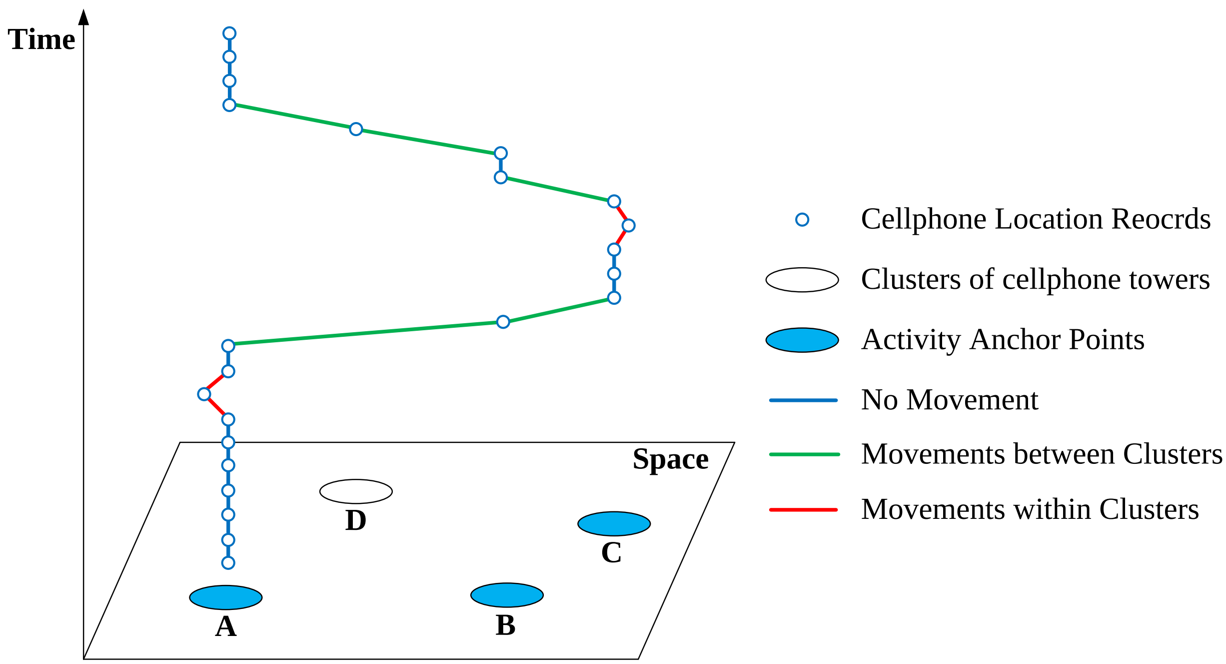



Figure 3.2 An individual's cellphone trajectory $T$ and key concepts in IMIs represented by a space-time system proposed by H ägerstrand (1970).

### *3.4.2 Estimation of individual home location*

Considering people's daily routines in most big cities in China (Long, Zhang and Cui 2012), we define home location of an individual as the activity anchor point with a minimum of four hours of stay at the location before 07:00. Based on this rule, we are able to estimate the home location for 97% of the sampled population in both Shanghai and Shenzhen. For each city, we compare our estimated home locations by administrative districts against the most recent census data (Gazette of the Sixth National Population Census for Shanghai Municipality 2010; Gazette of the People's Government of Shenzhen Municipality 2011). We find that our estimates are in agreement with the census data according to the population distribution by administrative districts (with Pearson coefficient at 0.95 and 0.99 for Shanghai and Shenzhen, respectively).

### *3.4.3 Building association rules*

Association rules have been widely used in business research to uncover the items that are frequently purchased together. They also have been used to describe associations between quantitative items or attributes (Han, Kamber and Pei 2011). To uncover the major characteristics of individual activity space in the two cities, one key challenge is to analyze how the three individual mobility indicators (IMIs) are associated with each other to characterize each individual's activity space. Note that the IMIs can be represented as:

$$IMIs \rightarrow (N, R, F) \tag{3.2}$$

where $N$ denotes the *number of activity anchor points*; $R$ describes an individual's *daily activity range*; and $F$ denotes the *frequency of movement*. The value of each indicator can be partitioned into several intervals:

$$N \rightarrow (N_1, N_2, \dots, N_a) \tag{3.3}$$

$$R \rightarrow (R_1, R_2, \dots, R_b) \tag{3.4}$$

$$F \rightarrow (F_1, F_2, \dots, F_c) \tag{3.5}$$

where $a$, $b$ and $c$ represent the number of intervals or classes defined for each corresponding indicator. For each individual $X$, the IMIs can be represented by its specific characteristics based on the defined intervals:

$$X \rightarrow (N_i, R_j, F_k) \qquad \text{(here } 0 \leq i \leq a, 0 \leq j \leq b, 0 \leq k \leq c) \tag{3.6}$$

We then introduce association rules to summarize the characteristics of human activity space for each city. The association rules are formulated as:

$$(X,\ "N_i") \Rightarrow (X,\ "\mathrm{R}_j"\ and\ "\mathrm{F}_k") \tag{3.7}$$

The above rules describe how different intervals of the three IMIs are associated with each other in each individual's activity space. For each city, the *support* and the *confidence* of the association rules are calculated as follows:

$$support(X,\ "N_i") \Rightarrow (X,\ "\mathrm{R}_j" and\ "\mathrm{F}_k") = \frac{\text{number of individuals with } (X,\ "N_i")}{\text{total population in the cellphone dataset}} \tag{3.8}$$

$$confidence(X,\ "N_i") \Rightarrow (X,\ "\mathrm{R}_j"\ and\ "\mathrm{F}_k") = \frac{\text{number of individuals with } (X,\ \ "N_i" \text{ and } "\mathrm{R}_j" \text{and } "\mathrm{F}_k")}{\text{number of individuals with } (X,\ "N_i")} \tag{3.9}$$

The *support* of a rule denotes the amount of individuals meeting the left-hand-side (LHS) condition divided by the total population of the dataset. The *confidence* of a rule denotes the amount of individuals meeting both sides of the rule divided by the number of individuals meeting the LHS condition. Both support and confidence indices describe important characteristics of human activity spaces extracted from a particular dataset. Note that we use $"N_i"$ as the LHS of the association rules because the *number of activity anchor points* for an individual $X$ is a discrete variable, which can be directly derived from individual cellphone trajectory data.

## 3.5 Analysis Results

### *3.5.1 General Statistics*

We first derive the general statistics of IMIs for the two cities. As shown in Figure 3.3a, the majority of people in Shenzhen had only 1 or 2 activity anchor points in the study day (38.8% and 38.5% of the population, respectively), while people in Shanghai were more diversified regarding the *number of activity anchor points* ($N$). For *daily activity range* ($R$), a large percentage of people in Shenzhen travelled within a very short distance during the day, as illustrated in Figure 3.3b. The cumulative distribution shows that nearly 50 percent of the people in Shenzhen travelled within $1.0km$ and about 82 percent travelled within $5.0km$. On the other hand, 26 percent of people in Shanghai travelled within $1.0km$ and 60 percent of people travelled within $5.0km$. The median of $R$ in Shenzhen and Shanghai are $1.1km$ and $3.1km$, respectively[3].

For *frequency of movement* ($F$), people on average made 3.76 movements (standard deviation: 3.50) in Shenzhen, as compared to 4.34 (standard deviation: 3.48) in Shanghai. The results indicate that: (1) people in Shanghai had more major activity locations (i.e., $N$) in a day than people in Shenzhen; (2) the spatial extent of people's activities in Shanghai were generally larger than that of people in Shenzhen; (3) people in Shanghai were more "active" in terms of the movements among their daily activity locations. However, it is still unclear how the three determinants ($N, R$ and $F$) are related to each other in an individual's activity space. For example, do people with the same number of activity anchor points in Shenzhen and Shanghai have similar daily activity range and/or movement frequency? In the next section, we discuss the

[3] The median of $R$ in Shenzhen and Shanghai ($1.1km$ and $3.1km$, respectively) are much lower than that of NY and LA regions ($6.08km$ and $8.0km$, respectively) computed by Isaacman et al. (2010) using cellphone location data. It is not surprising to see that the two U.S. cities have a larger daily activity range than Shanghai and Shenzhen since U.S. cities are more automobile oriented.

inter-relationships of $(N, R, F)$ based on association rules to further understand the differences and similarities of individual activity space in the two cities.

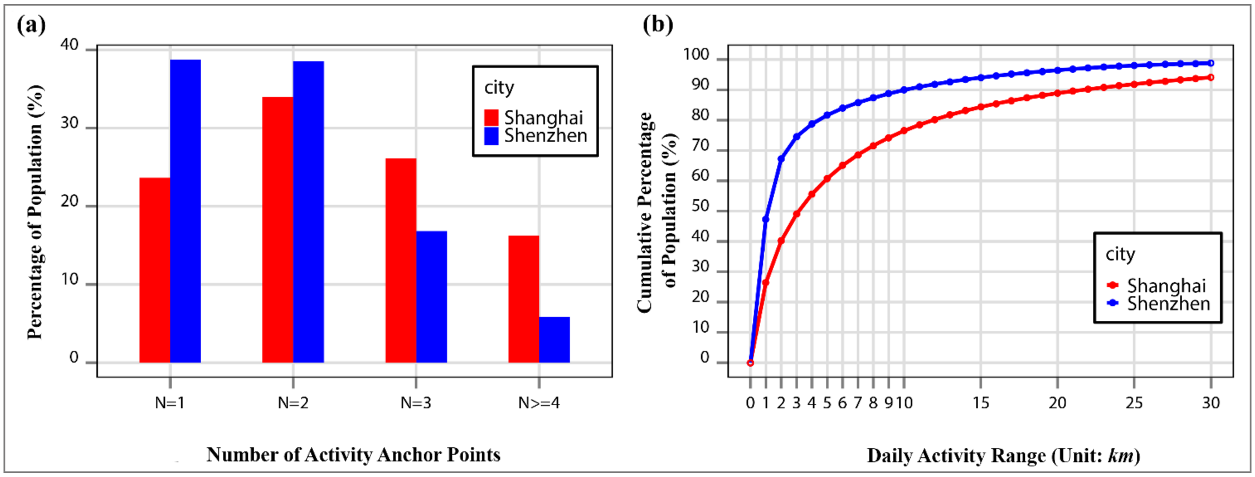



Figure 3.3 Distribution patterns of (a) number of activity anchor points, and (b) daily activity range in Shanghai and Shenzhen.

### *3.5.2 Association rules of IMIs*

To generate the association rules, we first partition the three individual mobility indicators into intervals. As shown in Table 3.2, we partition $N$, $R$ and $F$ into 4, 5 and 5 intervals, respectively. Each interval ($N_i$, $R_j$ or $F_k$) represents a particular value or range of values for the corresponding indicator. By mining the associations among the three indicators using the derived intervals, we are able to uncover the major characteristics of individual activity space for particular population groups within each city.

The support and confidence of the association rules are calculated to compare individual activity spaces in the two cities. As illustrated in Figure 3.4a$_1$ and 3.4b$_1$, although there are more people with one activity anchor point in Shenzhen ($support = 38.8\%$) than in Shanghai

($support = 23.6\%$), people in these two subsets ($N = 1$) had very similar activity space characteristics. The two subsets are dominated by individuals with very short daily activity range ($R_1$) and low movement frequency ($F_1$ and $F_2$). Only a very small percentage of people travelled very far ($R_4$ and $R_5$) and frequently ($F_4$ and $F_5$).

Table 3.2 Intervals (classes) defined for the association rules.

| **Number of activity anchor points ($N$)** | | **Daily activity range ($R$)** | | **Frequency of movement ($F$)** | |
|---|---|---|---|---|---|
| Intervals | Values | Intervals | Values | Intervals | Values |
| $N_1$ | $N = 1$ | $R_1$ | $R \leq 1km$ | $F_1$ | $F = 0$ |
| $N_2$ | $N = 2$ | $R_2$ | $1km < R \leq 2km$ | $F_2$ | $1 \leq F \leq 3$ |
| $N_3$ | $N = 3$ | $R_3$ | $2km < R \leq 5km$ | $F_3$ | $4 \leq F \leq 7$ |
| $N_4$ | $N \geq 4$ | $R_4$ | $5km < R \leq 10km$ | $F_4$ | $8 \leq F \leq 11$ |
| | | $R_5$ | $R > 10km$ | $F_5$ | $F \geq 12$ |

When $N = 2$ (Figure 3.4a$_2$ and 3.4b$_2$), the percentages of people with different travel ranges distribute relatively evenly within each interval of $R$ in Shanghai, while Shenzhen shows a decay with increasing travel range. The subset of Shenzhen are dominated by people with short daily activity ranges (e.g., 63.2 percent of people with $R \leq 2km$). In the subset of Shanghai, however, more people tended to travel further in a day (e.g., 36 percent with $R > 5km$). The observed difference could be potentially explained by the home-work relationships of people in the two subsets considering that home and workplace are two primary activity locations for most people. How frequently people moved serves as an important indicator of urban dynamics. According to our observation, although the majority of people in both subsets fall within $F_2$ and

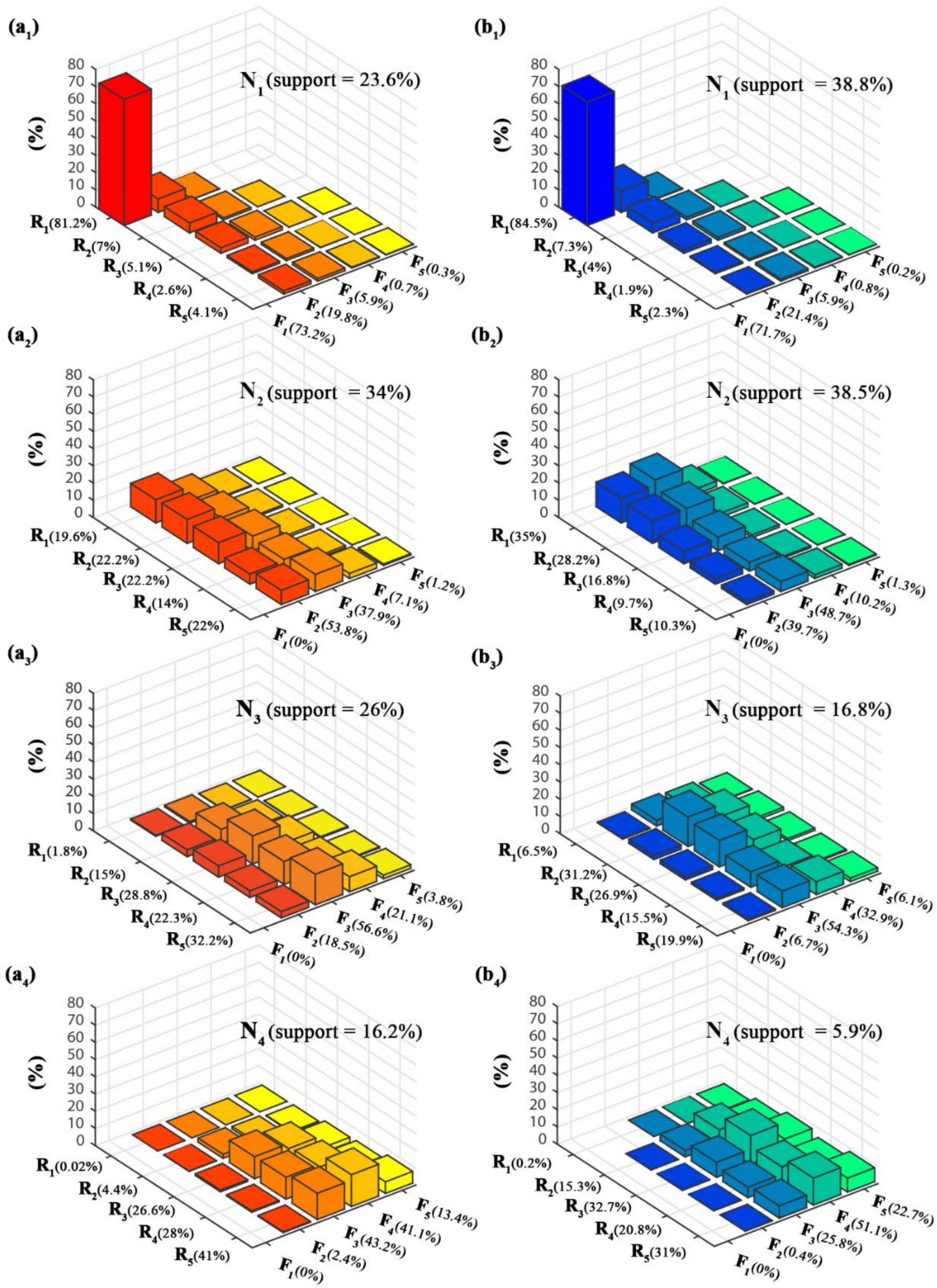



Figure 3.4 Association rules of individual activity space: ($a_1$ to $a_4$) confidence of rules with left-hand-side (LHS) organized by $N_i$ in Shanghai; ($b_1$ to $b_4$) confidence of rules with LHS organized by $N_i$ in Shenzhen. For each individual figure, the percentages next to interval labels denote the sum of confidence for the corresponding rows or columns.

$F_3$, people in Shenzhen travelled more frequently (39.7% and 48.7% within $F_2$ and $F_3$) than people in Shanghai (53.8% and 37.9% within $F_2$ and $F_3$). Note that we analyze the temporal variations of people's movement patterns in the two cities later in this section to further examine when (and where) people were more active in their daily activity spaces.

There is a notable change of the distribution of association rules as $N$ increases from 2 to 3 for people in Shanghai. The majority of people with $N = 3$ (Figure 3.4a$_3$) in Shanghai had a very large daily activity range (83.3% with $R > 5km$), which is quite different from the relatively even distribution of $R$ when $N = 2$ (Figure 3.4a$_2$). The result indicates that the third activity anchor point might have a significant effect on people's travel range in Shanghai. Shenzhen, however, still has a large proportion of people with a short daily activity range (64.6% with $R \leq 5km$, as shown in Figure 3.4b$_3$), which is similar to what we observe when $N = 2$ (Figure 3.4b$_2$).

According to the comparisons among the three subgroups in the two cities, we can see that in Shenzhen, a small activity space was usually enough to fulfill various purposes of people's daily activities such as working, dining, recreation, and so forth. In Shanghai, activity locations were more widely distributed in an individual's activity space. People were more likely to travel far from their primary activity locations (i.e., home and workplace) for certain travel/activity purposes. For $N \geq 4$ (Figure 3.4b$_4$ and 3.4b$_4$), we see an increase in both travel range and movement frequency in both cities as compared to the previous three subgroups. People in Shanghai still travelled further but less frequently, as compared to the same population group in Shenzhen. Note that we've tested other partition schemes to generate different intervals for IMIs and the derived association rules reveal similar patterns of people's activity spaces in the two cities.

### *3.5.3 Spatial variations of human activity space*

Analyzing the geographic patterns of people's activity space within the context of built environment could produce an improved understanding of their daily activity patterns. For example, it would be meaningful to explore the geographic distributions of people with a small daily activity range ($R$), which is an important feature of individual activity space in both cities, especially in Shenzhen. Figure 3.5 illustrates the geographic distributions of people with a daily activity range $R \leq 2km$. Specifically, we divide the study areas into $2km$ grids, and aggregate individuals based on their estimated home locations. Each grid cell represents the number of individuals with $R \leq 2km$, normalized by the total number of individuals in that grid cell.

As shown in Figure 3.5a, many grid cells in the core areas of Shanghai have a higher percentage of people with $R \leq 2km$ (i.e., green cells in Huangpu, Luwan and Jingan districts, readers can refer to the inset map in Figure 3.1b) as compared to the grid cells in suburbs (i.e., orange and red cells) such as Jinshan, Songjiang, Qingpu, Jiading and Pudongxinqu districts. Note that we also observe grid cells with higher percentages (i.e., green cells) in certain suburbs such as Minxing and Fengxian districts. It is interesting to find that the observed patterns are in general agreement with the analysis results by Sun, Pan and Ning (2008) who studied job-housing balance in Shanghai. They indicate that core areas such as Huangpu, Luwan and Jingan have more job opportunities as compared to the number of residents, thus more people would have a relatively shorter commuting distance. Some suburbs around the core areas are more housing-oriented, thus more people would have a longer commuting distance. We avoid making any further statement since the daily activity range examined in this study does not reflect people's actual commuting distances.

Figure 3.5b shows that there is a general "north-south" divide in Shenzhen and the proportion of people with $R \leq 2km$ in most grid cells of the two northern districts (Baoan and Longgang) are larger than 60% or even 80%, which indicates that most people who live around these cells have a small daily activity range during the study day. In the southern part of Shenzhen, the percentages are generally lower. To explore potential causes to the identified patterns, we also display the locations of major factories in Shenzhen. It appears that grid cells with a high percentage are generally co-located with major factories in Shenzhen. Many factories in Shenzhen provide workers with dormitories adjacent to their workplace. In addition, many immigrants tend to rent apartments near their workplace to save commuting time and cost. The findings suggest that the geographic patterns of people's activity space in Shenzhen and in Shanghai are quite different, and the identified patterns are likely to be related to the underlying socio-economic characteristics.

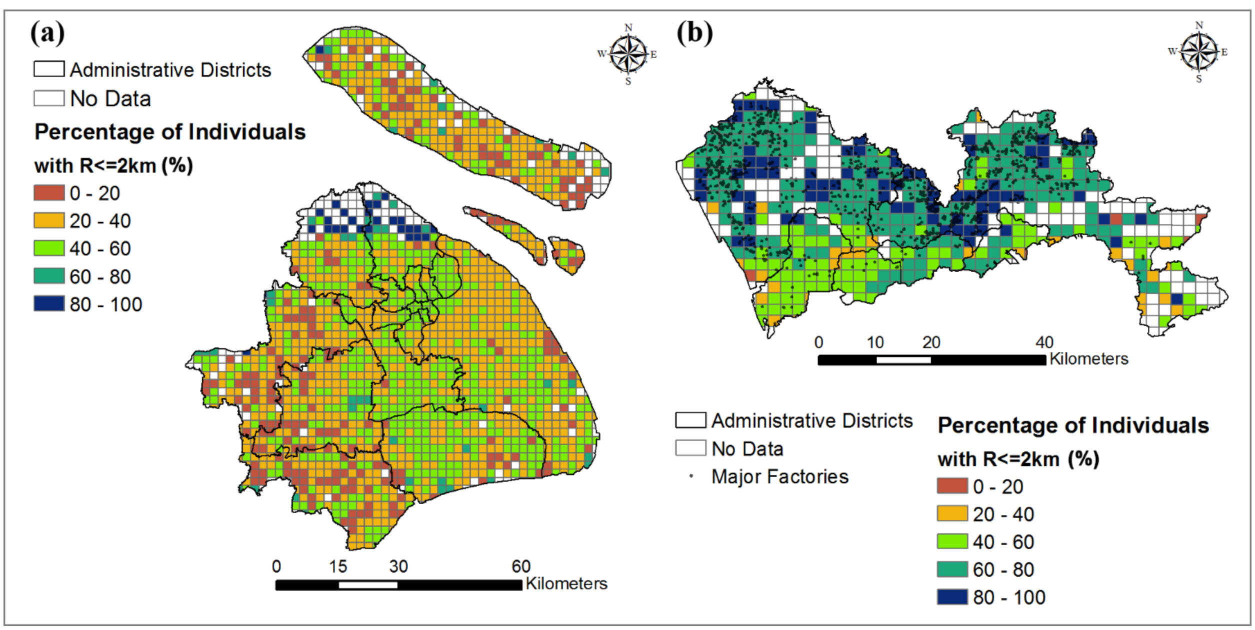



Figure 3.5 Geographic distributions of individuals with daily activity range $\leq 2km$ in the two cities: (a) geographic patterns in Shanghai; (b) geographic patterns in Shenzhen.

### *3.5.4 Temporal variations of aggregate movement patterns*

We further analyze the temporal variations of aggregate movement patterns to understand when people were more active in their daily activity spaces. Figure 3.6 shows the percentages of people who moved through a day in the two cities, organized by the number of individual activity anchor points ($N$). People with $N = 1$ in the two cities didn't move much in a day. The percentages are relatively stable over time (less than 10%). As expected, movement patterns of the other three subgroups in these two cities exhibit two peaks during the morning and afternoon rush hours. However, there is a local peak around time intervals 12 and 13 for people in Shenzhen, which indicates that people in Shenzhen moved more frequently around noon than other work hours. More importantly, the difference of aggregate movement patterns around noon between the two cities explains our previous finding that people in Shenzhen generally move more frequently than people in Shanghai when controlling the number of activity anchor point ($N$).

We further explore the temporal variations of average movement distances in the two cities. As shown in Figure 3.7, people's movement distances in both cities are generally lower around noon than the work hours. By further analyzing movements around noon (i.e., time intervals 12 and 13), we find that 43% of individuals travel from or to their home locations around noon in Shanghai, as compared to 66% in Shenzhen. The shorter movement distance around noon reveals an interesting aspect of people's lifestyle in both cities. Although people in Shanghai have a longer travel distance in general, short range movements still dominate in both cities. As described previously, the share of non-motorized trips accounts for more than 50% of all trips in both Shanghai and Shenzhen. Our analysis results suggest that travel mode such as walking and bicycle should receive more attention in urban and transportation planning that have

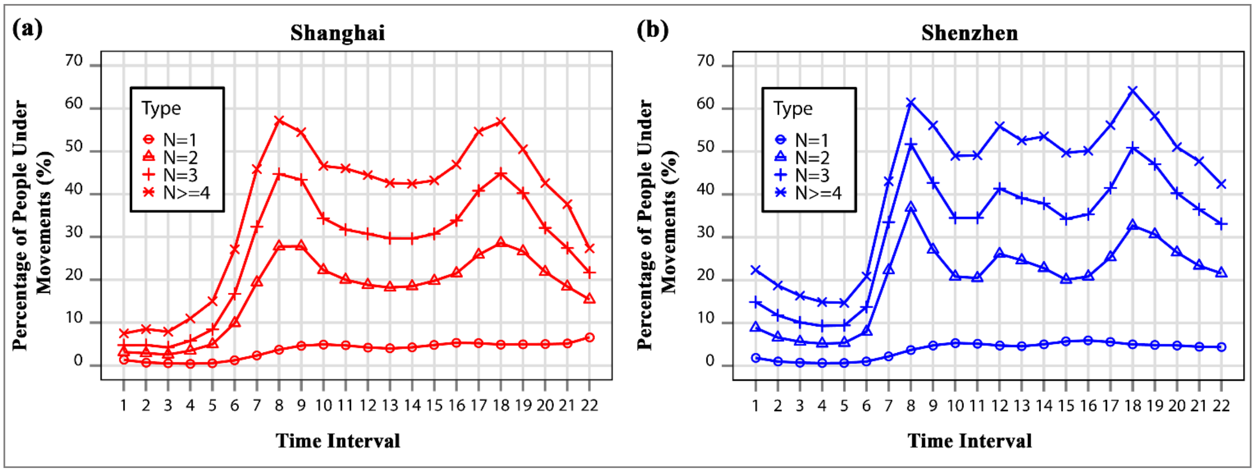



Figure 3.6 Temporal variations of aggregate movement patterns in the two cities, organized by the number of individual activity anchor points: (a) temporal patterns in Shanghai, and (b) temporal patterns in Shenzhen. (Each time interval is associated with two consecutive time windows in Table 3.1. For example, time interval 1 in this figure shows movement patterns from 00:00-01:00 to 01:00-02:00 in Table 3.1).

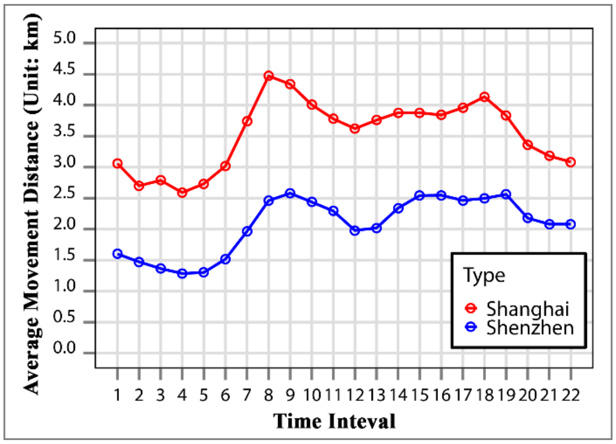



Figure 3.7 Temporal variation of the average movement distance in Shanghai and Shenzhen.

been mentioned in many government reports in recent years (e.g., Urban Planning Land and Resources Commission of Shenzhen Municipality, 2013). Appropriate transportation services should be deployed to accommodate medium- and short-range trips in cities such as Shenzhen and Shanghai.

### 3.6 Discussion and Conclusion

The emergence of big individual tracking datasets brings new opportunities and challenges to the understanding of human activity space in our urban environment. In this study, we develop several intuitive individual mobility indicators (IMIs) using cellphone location data, to describe the major determinants of individual activity space. Different from previous studies (Kang et al. 2010; Isaacman et al. 2012; Becker et al. 2013) which investigate determinants of human activity space independently, we analyze how these mobility indicators (e.g., daily activity range, number of activity anchor points, and frequency of movement) combine with each other in an individual's activity space. The association rules of IMIs are able to uncover the complexities of individual activity space for a given population or a geographic region. The support and confidence of the derived rules serve as a signature of people's daily activity patterns, and enable us to compare human activity spaces systematically across different geographic regions. By using active tracking cellphone location datasets collected in Shanghai and Shenzhen, we summarize and compare the major characteristics of activity space patterns in these two cities. The association rules and spatiotemporal analysis of aggregate human activity patterns allow us to better understand the socioeconomic characteristics of these cities, and yield some insights into transportation planning and urban design.

Our analysis results reveal several interesting aspects of human activity spaces in the two cities, and the implications are worth discussing. First, quite a few people in both cities stayed around one particular location for the whole day. Such unique activity patterns might reflect some societal issues such as "urban villages" (Wei and Yan 2005) in cities that consist of low-income communities of migrant population. Additional efforts are needed to further examine the "immobility" of these people and the potential driving forces related to land use planning (Pan, Shen and Zhang 2009) and social segregation (Schönfelder and Axhausen 2003; Silm and Ahas 2014). Second, for the majority of people in Shenzhen, a small activity space was usually enough to fulfill the needs of people's daily activities, which is consistent with the government's goal of building a compact city with sustainable urban form. In Shanghai, however, activity locations were more widely distributed in an individual's activity space, and people were more likely to travel far from their home and workplaces for certain activity purposes. Shenzhen and Shanghai, one being a city with large migrant population and the other being a century-old metropolis with many local residents, have very different socio-demographic characteristics and urban forms, which play an important role in shaping people's daily activity patterns. Third, the geographic disparity of people's travel range in Shenzhen is significant. The difference between the north and south could be partially explained by the socio-economic divide in Shenzhen. In Shanghai, the geographic disparity is less obvious, and our analysis suggests that the identified patterns could be potentially explained by people's commuting patterns and the job-housing relationships in the city.

Currently, the research findings only reflect people's activity space in the two cities for a day. In the future, we plan to further investigate the temporal variations of individual activity space (e.g., seasonality, and difference between workdays and weekends) by using active tracking cellphone datasets that cover longer time periods. It would also be meaningful to compare the analysis results derived from active and passive cellphone location data (e.g., CDRs), for example, to examine whether they reveal similar or different patterns of people's activity space. This will help us better understand the strength and weakness of each data type, and the intrinsic characteristics of human activity space. Nevertheless, the research findings in this article enhance our understanding of the geographies of human mobility in a space-time context. We believe the proposed methods are useful to other types of large individual tracking datasets for data-intensive analyses of human activity space.

## References


Ahas, R., S. Silm, O. Järv, E. Saluveer, and M. Tiru. 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology* 17 (1):3-27.

Axhausen, K. W., A. Zimmermann, S. Schönfelder, G. Rindsfüser, and T. Haupt. 2002. Observing the rhythms of daily life: A six-week travel diary. *Transportation* 29 (2):95-124.

Becker, R., R. Cáceres, K. Hanson, S. Isaacman, J. M. Loh, M. Martonosi, J. Rowland, S. Urbanek, A. Varshavsky, and C. Volinsky. 2013. Human mobility characterization from cellular network data. *Communications of the ACM* 56 (1):74-82.

Brdar, S., D. Culibrk, and V. Crnojevic. 2012. Demographic attributes prediction on the real-world mobile data. Nokia mobile data challenge 2012 workshop, Newcastle, UK.

Brown, L. A., and E. G. Moore. 1970. The intra-urban migration process: a perspective. *Geografiska Annaler. Series B, Human Geography* 52 (1):1-13.

Buliung, R. N., and P. S. Kanaroglou. 2006. A GIS toolkit for exploring geographies of household activity/travel behavior. *Journal of Transport Geography* 14 (1):35-51.

Chen, J., S.-L. Shaw, H. Yu, F. Lu, Y. Chai, and Q. Jia. 2011. Exploratory data analysis of activity diary data: a space–time GIS approach. *Journal of Transport Geography* 19 (3):394-404.

Cheng, Z., J. Caverlee, K. Lee, and D. Z. Sui. 2011. Exploring Millions of Footprints in Location Sharing Services. ICWSM 2011:81-88.

Candia, J., M. C. González, P. Wang, T. Schoenharl, G. Madey, and A.-L. Barabási. 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41 (22):224015.

Cho, E., S. A. Myers, and J. Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 1082-1090.

Dijst, M. 1999. Two-earner families and their action spaces: A case study of two Dutch communities. *GeoJournal* 48 (3):195-206.

Gao, H., J. Tang, and H. Liu. 2012. Mobile location prediction in spatio-temporal context. Nokia mobile data challenge 2012 workshop, Newcastle, UK.

Gazette of the People's Government of Shenzhen Municipality (in Chinese). 2011. Issue No. 17, Serial No. 741. http://www.sz.gov.cn/zfgb/2012_1/gb785/201204/t20120423_18 44697.htm (last accessed 28 October 2014)

Gazette of the Sixth National Population Census for Shanghai Municipality (in Chinese). 2010. http://www.stats-sh.gov.cn/sjfb/201105/218819.html (last accessed 28 October 2014)

Golledge, R. G., and Stimson, R. J. 1997. Spatial behavior: A geographic perspective: Guilford Press.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196):779-782.

Hägerstrand, T. 1970. What about people in regional science? *Papers in Regional Science* 24 (1):7-24.

Han, J., M. Kamber, and J. Pei. 2006. Data mining: concepts and techniques, 3rd edition. Elsevier: Morgan kaufmann.

Hanson, S. 1980. The importance of the multi-purpose journey to work in urban travel behavior. *Transportation* 9 (3):229-248.

Hanson, S., and P. Hanson. 1981. The travel-activity patterns of urban residents: dimensions and relationships to sociodemographic characteristics. *Economic geography*:332-347.

Horton, F. E., and D. R. Reynolds. 1971. Effects of urban spatial structure on individual behavior. *Economic geography*:36-48.

Isaacman, S., R. Becker, R. Cáceres, S. Kobourov, J. Rowland, and A. Varshavsky. 2010. A tale of two cities. Proceedings of the Eleventh Workshop on Mobile Computing Systems & Applications, pp. 50-51.

Isaacman, S., R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. 2012. Human mobility modeling at metropolitan scales. Proceedings of the 10th international conference on Mobile systems, applications, and services.

Kang, C., S. Gao, X. Lin, Y. Xiao, Y. Yuan, Y. Liu, and X. Ma. 2010. Analyzing and geo-visualizing individual human mobility patterns using mobile call records. 18th International Conference on Geoinformatics: 1-7.

Kwan, M. P. 1999. Gender, the Home-Work Link, and Space-Time Patterns of Nonemployment Activities*. *Economic geography* 75 (4):370-394.

Lenntorp, B. 1977. Paths in space-time environments: A time-geographic study of movement possibilities of individuals. *Environment and Planning A* 9 (8):961-972.

Li, Q., Y. Zheng, X. Xie, Y. Chen, W. Liu, and W.-Y. Ma. 2008. Mining user similarity based on location history. Proceedings of the 16th ACM SIGSPATIAL international conference on Advances in geographic information systems.

Long, Y., Y. Zhang and C. Cui. 2012. 利用公交刷卡数据分析北京职住关系和通勤出行 (Identifying Commuting Pattern of Beijing Using Bus Smart Card Data). *Journal of Geographical Sciences* 67(10): 1339-1352.

Lynch, K. 1960. The image of the city: MIT press.

National Bureau of Statistics of China. 2012. 主要城市年度GDP (Annual GDP for Major Cities in China). http://data.stats.gov.cn/workspace/index?m=csnd (last accessed 28 October 2014)

Newsome, T. H., W. A. Walcott, and P. D. Smith. 1998. Urban activity spaces: Illustrations and application of a conceptual model for integrating the time and space dimensions. *Transportation* 25 (4):357-377.

Pan, H., Q. Shen, and M. Zhang. 2009. Influence of urban form on travel behaviour in four neighbourhoods of Shanghai. *Urban studies* 46 (2):275-294.

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding*, 14-25: Springer.

Schönfelder, S., and K. W. Axhausen. 2003. Activity spaces: measures of social exclusion? *Transport Policy* 10 (4):273-286.

Shaw, S. L., H. Yu, and L. S. Bombom. 2008. A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12 (4):425-441.

Shen, Y., M.-P. Kwan, and Y. Chai. 2013. Investigating commuting flexibility with GPS data and 3D geovisualization: a case study of Beijing, China. *Journal of Transport Geography* 32:1-11

Shenzhen Daily. 2012. "Shenzhen: Most crowded in China". http://szdaily.sznews.com/html/2012-05/30/content_2063502.htm (last accessed 03 October 2015).

Shoval, N. and M. Isaacson. 2007. Sequence alignment as a method for human activity analysis in space and time. *Annals of the Association of American Geographers* 97(2):282-297.

Silm, S. and R. Ahas. 2014. Ethnic Differences in Activity Spaces: A Study of Out of-Home Nonemployment Activities with Mobile Phone Data. *Annals of the Association of American Geographers* 104(3):542-559.

Song, C., Z. Qu, N. Blumm, and A.-L. Barabási. 2010. Limits of predictability in human mobility. *Science* 327 (5968):1018-1021.

Wei, L., and X. Yan. 2005. Transformation of "Urban Village" And Feasible Mode [J]. *City Planning Review* 7:9-13.

Xu, Y., S.-L. Shaw, Z. Zhao, L. Yin, Z. Fang, and Q. Li. 2015. Understanding aggregate human mobility patterns using passive mobile phone location data: a home-based approach. *Transportation* 42(4):625-646.

Yuill, R. S. 1971. The standard deviational ellipse; an updated tool for spatial description. *Geografiska Annaler. Series B. Human Geography*:28-39.

Zheng, Y., L. Liu, L. Wang, and X. Xie. 2008. Learning transportation mode from raw gps data for geographic applications on the web. Proceedings of the 17th international conference on World Wide Web.

Zhu, Y., E. Zhong, Z. Lu, and Q. Yang. 2012. Feature engineering for place category classification. Nokia mobile data challenge 2012 workshop, Newcastle, UK.

# CHAPTER 4

# UNCOVER POTENTIAL DEMAND OF BICYCLE TRIPS USING CELLPHONE LOCATION DATA – AN ANCHOR-POINT BASED TRAJECTORY SEGMENTATION METHOD

This chapter is a manuscript in preparation for *Journal of Transport Geography*.

**Abstract:** A growing number of cities around the world are promoting bike sharing systems to mitigate urban problems related to public health, traffic congestion, energy consumption and air pollution. In order to build a successful bike sharing system, knowing where the demands are and when they occur is of primary importance. This study uses a large cellphone location dataset to uncover potential demand of bicycle trips in a city. By identifying two important anchor points (night-time and day-time anchor points) from individuals, this study introduces an anchor-point based trajectory segmentation method to partition cellphone trajectories into four types of trip chain segments. By selecting trip chain segments that can be potentially served by bicycles, two indicators ($inflow_p$ and $outflow_p$) are generated at the cellphone tower level to estimate potential demand of incoming and outgoing bicycle trips at different places in a city and times of a day. A location-allocation model is performed to suggest locations of bike sharing stations based on the total demand generated at each cellphone tower. Two measures (accessibility and dynamic relationships between incoming and outgoing trips) are introduced to understand the characteristics of the bike stations. The accessibility measure quantifies how well the stations could serve bicycle users to reach other potential activity destinations. The dynamic relationships (between the incoming and outgoing trips) reflect the asymmetry of human travel patterns at different times of a day, which serve as useful information for the distribution and redistribution of bicycles among the bike stations.

**Keywords:** Mobile phone location data, anchor point, trajectory segmentation, bike station location.

## 4.1 Introduction

Public bicycle programs have received increasing attention in the past few decades. Many cities around the world are promoting bicycle use to mitigate urban problems related to public health, traffic congestion, energy consumption and air pollution. Bike sharing systems have been implemented in many cities to offer short-term bike rental services to individuals for point-to-point trips. A successful bike sharing system could encourage people's use of bikes for short distance trips and ease traffic pressure in congested urban areas. However, it is not an easy task to decide where investments should be made to deploy bicycle infrastructures (e.g., bike sharing stations) that accommodate people's travel needs. Among various factors that need to be considered, knowing where the demands are and when they occur is of primary importance.

Travel surveys and census data have been widely used in previous studies (Cervero and Dunkan 2003; Barnes and Krizek 2005; Larsen, Patterson and El-Geneidy 2013) to estimate or model the demand of bicycle usage, and provide decision support for locating new cycling facilities such as bike sharing stations. However, collecting travel surveys and census data is usually expensive and time-consuming. Moreover, the amount of information that can be collected is largely restricted due to limited human and financial resources. Recent advancements in location-aware technologies have provided many new data sources (e.g., smart card data, GPS tracking data, and mobile phone location data) for understanding whereabouts of people in space and time. These large-scale anonymized datasets introduce new opportunities for researchers and planners to estimate travel demand related to various transportation modes in a city. However, there has been limited research on how to estimate potential demand of bicycle trips using these newly emerged data sources to facilitate the planning of bike sharing stations in a city.

In recent years, mobile phone location data have been widely used to study human mobility patterns and people's use of urban space. Among these studies, considerable efforts have been devoted to uncovering people's major activity locations (e.g., home and workplace) as well as movement patterns among these locations (Candia et al. 2008; Ahas et al. 2010; Cho, Myers and Leskovec 2011; Becker et al. 2013). Such information reflects how people plan their daily travels among important activity destinations and describes people's daily trip chains (Strathman and Dueker 1995; McGuckin, Zmud and Nakamoto 2005; Golob and Hensher 2007). These activity destinations and daily trip chains serve as important components of people's daily travels and can be used to estimate travel demand related to various transportation modes (e.g., cycling) in a city. Hence, this study uses a large scale mobile phone location data collected in Shenzhen, China to uncover potential demand of bicycle trips in the city. The main contributions of this research are as follows:

(1) By identifying two important anchor points (night-time and day-time anchor points) from individual cellphone trajectories, this study introduces an anchor-point based trajectory segmentation method to partition the cellphone trajectories into meaningful trip chain segments. By selecting trip chain segments that can be potentially served by bicycles, two indicators ($inflow_p$ and $outflow_p$) are generated at the cellphone tower level to estimate potential demand of incoming and outgoing bicycle trips at different places in the city and times of a day. The two indicators reflect the intensity and daily rhythms of people's short distance trips at a relatively fine spatial resolution, and can be further used to assist planning of bicycle infrastructures such as bike sharing stations in the city.

(2) A location-allocation model is performed to suggest locations of bike sharing stations in different scenarios (e.g., 300, 600, 900 and 1200 bike stations) based on the potential demand

generated at each cellphone tower. Two measures (accessibility and dynamic relationships between incoming and outgoing trips) are introduced to understand the characteristics of bike sharing stations once their locations are derived. The accessibility measure quantifies how well the stations could serve bicycle users to reach other potential activity destinations. The dynamic relationships between incoming and outgoing trips reflect the asymmetry of human travel patterns at different times of a day, which serve as useful information for the distribution and redistribution of bicycles among the bike stations in the city.

## 4.2 Literature Review

### *4.2.1 Bike sharing systems*

Bike sharing systems have received growing attention in recent years with initiatives to reduce traffic congestion, increase accessibility, and improve the health of residents (DeMaio and Gifford 2004). According to a report (The Bike-Share Planning Guide) provided by the Institute for Transportation & Development Policy (ITDP) in 2013, more than 600 cities[1] around the world have established their own bike sharing systems and more are starting every year. The evolution of bike sharing systems over the past 50 years can be categorized into three generations (DeMaio 2003; Shasheen, Guaman and Zhang 2010). The first generation of bike-sharing systems, also known as Free Bike System, was released in Amsterdam in 1965. The system was provided for public use at no charge and the bicycles were unlocked so that users could drop them off at any place they wanted. However, the bike-sharing system suffered from problems such as theft and vandalism, and collapsed within a short period of time. The second

[1] Examples of these bike sharing systems are *Vélib* in Paris, Fance (http://www.velib.paris/), *Bicing* in Barcelona, Spain (https://www.bicing.cat/), *Call-a-Bike* in Germany (http://www.callabike.de/), *Cycle Hire* in London, United Kingdom, and *Ecobici* in Mexico City, Mexico (https://www.ecobici.df.gob.mx/).

generation of bike sharing systems, known as the coin-deposit system, was first established in Nakskov, Denmark in 1993, followed by a larger bike-sharing program launched in Copenhagen in 1995. Users could pick up and return the bicycles at specific locations using a coin deposit. The third generation of bike-sharing systems, known as the information technology-based system, was first introduced in England in 1996. The third generation of bike-sharing systems incorporated many new technologies at that time such as smartcards, mag-stripe cards and mobile-phone access (Wang et al. 2009). DeMiao (2009) and Shasheen, Guzman and Zhang (2010) also gave an outlook to the fourth generation of the bike sharing system, which incorporates more advanced technologies such as improved distribution, ease of installation, tracking, pedal assistance, and anti-theft mechanism.

### *4.2.2 Forecasting bicycle travel demand*

In order to establish a successful bike sharing system, planners need to have a good idea about where the travel demands are as well as other factors related to land topography, connectivity of transportation networks, land-use diversity, weather and safety (Cervero and Duncan 2003; Harkey, Reinfurt and Knuiman 1998; Iacono, Krizek and El-Geneidy 2010). According to Porter, Suhrbier and Schwartz (1999), previous studies usually adopted four broad categories of methods to estimate demand of bicycle trips, which are aggregate-level methods, attitudinal surveys, discrete choice models and regional travel models (e.g., “four-step travel demand models”). Most of these methods relied on detailed information about human daily activities (e.g., surveys) and/or numerous assumptions about human travel behavior (e.g., discrete choice models). For example, Landis (1996) proposed a Latent Demand Score (LDS) model based on a probabilistic gravity model to estimate the amount of bicycle trips that would occur on a road segment. Clark (1997) used a four-step travel demand model to estimate the length and travel time of the trips in Bend, Oregon to identify travels that could be made by bicycles. Rybarczyk and Wu (2010) introduced the bicycle level of service index and demand potential index to analyze the spatial relationships between bicycle supply and demand. The demand of bicycle trips was estimated based on several factors related to population distribution as well as locations of parks, recreation areas, schools and businesses. Wardman, Tight and Page (2007) developed a mode choice model which combined revealed preference data (with individual's actual mode choices) and stated preference data (with hypotheses on individual choices among different alternatives) to predict the future trends in commuter cycling in Great Britain. Although travel surveys and regional travel demand models are valuable for the estimation of bicycle trips, it usually requires a mass of human resources and financial input to collect the data (i.e. surveys). Moreover, many travel demand models used "zone structures that are too large to be of much use in deciding on the size and location of bike-share stations" (ITDP 2013, pp. 56).

*4.2.3 Mobile phone location data for travel behavioral analysis*

Recent advancements of location-aware technologies have produced many new data sources (e.g., cellphone location data and social media data) for understanding whereabouts of people in space and time. These new datasets enable the studies of human activities "at low cost and on an unprecedented scale" (Becker et al. 2013, pp. 74). For example, many studies have used cellphone location data to solve problems related to the characterization and prediction of human mobility patterns (Candia et al. 2008; Gonzalez, Hidalgo and Barabasi 2008; Bayir, Demirbas and Eagle 2010; Phithakkitnukoon er al. 2010; Song et al. 2010; Cho, Myers and Leskovec 2011; Yuan, Raubal and Liu 2012; Becker et al. 2013; Calabrese et al. 2013; Csáji et al. 2013; de Montjoye et al. 2013), and various aspects of urban dynamics (Ahas et al. 2007;

Ratti et al. 2007; Reades, Calabrese and Ratti 2009; Ahas et al. 2010; Vieira et al. 2010;). Among these studies, considerable effort has been devoted to uncovering people's use of urban space and daily rhythms of urban flows. However, there has been limited research on how to better understand potential demand of bicycle trips from cellphone location data.

In the past few years, there have been some studies which used cellphone location data to study individual trip making, especially for the trips that are tied to people's key activity locations (e.g., home and workplace). For example, Iqbal et al. (2014) used Call Detail Record (CDRs) in Dhaka, Bangladesh to generate tower-to-tower transient OD matrices, which were then associated with traffic network and converted to node-to-node transient OD matrices. Similarly, Alexander et al. (2015) used CDRs in Boston metropolitan area over a period of two months to estimate origin-destination trips by purpose (e.g., home-based work trips, home-based other trips and non home-based trips). Dong et al. (2015) used CDRs to suggest traffic zone division in urban areas to assist travel demand forecast. Wang et al. (2012) used mobile phone location data collected in San Francisco and Boston area to evaluate urban road usage patterns. It is clear that mobile phone location data can be leveraged to understand people's travel demand related to various transportation modes and different types of human activities.

### *4.2.4 Bike stations and location-allocation models*

One of the most important tasks in planning a bike-sharing system is to determine the location of bike stations. Choosing good locations of bike stations would ensure that the system meets the current demand and even stimulates people's use of bicycles in the future. Many studies have given their thoughts to where bike stations should be located under particular application or research contexts. For example, Larsen, Patterson and El-Geneidy (2013) proposed a prioritization index calculated at the grid-cell level to demonstrate how to prioritize cycling infrastructure investments. The prioritization index was aggregated from several indicators including origin destination (OD) of actual bicycle trips, OD of short car trips, cyclists' route preferences, and concentration of bicycle crashes. Martinez et al. (2012) proposed a heuristic algorithm which encompassed a mixed integer linear program (MILP) and a p-median location-allocation problem to optimize the location of bike sharing stations in Lisbon, Portugal. The locations of bike stations were determined based on a list of factors related to user demand, the required investment and operational costs. García-Palomares, Gutiérrez and Latorre (2012) used the population and number of jobs at the building level along with the mobility survey at the transport zone level to estimate potential demand of bicycle trips in central Madrid. The authors adopted two location-allocation models with different objective functions (minimize impedance and maximize coverage) to suggest facility locations of bike sharing stations.

Some studies above adopted location-allocation models in order to suggest the optimal locations of bike stations with relation to the distribution of potential demand. According to Cooper (1961) and Rushton (1979), the location-allocation models aim at determining the number and/or locations of facilities (e.g., distribution centers) to meet some predefined objectives while satisfying the requirements at the demand points. The location-allocation models could vary depending on the objectives to be achieved. For example, the p-median problem (Hakimi 1965) and the p-center problem (Hakimi 1965; Suzuki and Drezner 1996) are two typical forms of location-allocation problems. The objective of the p-median problem is to locate p facilities to minimize the total weighted travel cost from the demand points to the facilities. The p-center problem aims at providing p facilities to minimize the maximum distance from a demand point to its closet facility. Toregas, Swain, Revelle and Bergman (1971) introduced the Location Set Covering Problem with the objective to determine the minimal number and location of facilities so that all demand points will fall within the maximal service distance from a facility. Based on this model, Church and Revelle (1974) formulated the Maximal Covering Location Problem (MCLP), which maximizes the population (or demand) within the service distance (of the facilities) by locating a fixed number of facilities.

In this study, the ArcGIS software (i.e. ArcGIS 10.1) by the Environmental Systems Research Institute (ESRI) and its Location-allocation module (included in the Network Analyst toolbox) are used to perform the location-allocation analysis to suggest the locations of bike sharing stations.

## 4.3 Study Area and Dataset

Shenzhen is a major city in Guangdong Province, China (Figure 4.1a). The city situates north of Hong Kong (Figure 4.1b) and is part of the so-called Pearl River Delta mega city (Telegraph, 2011). It is now a major finance and technology center in southern China, with an estimated population of 15 million as of 2012 (Shenzhen Daily 2012) and a total area of 2,050 $km^2$. As shown in Figure 4.1c, the city has 6 administrative districts and four management new districts[2] (List of administrative divisions of Shenzhen 2015). Shenzhen was a small fishing village when it became China's first Special Economic Zone (SEZ) in 1979. The SEZ comprised only Nanshan, Futian, Luohu and Yantian districts until 1st July 2010, and was then expanded to include all the other districts (Shenzhen Special Economic Zone 2015). The southern and northern parts of Shenzhen have different socioeconomic and demographic characteristics. The four districts in the southern part of Shenzhen (i.e., Nanshan, Futian, Luohu and Yantian) are

[2] Guangming and Longgang are two management new districts that are subordinate to Bao'an district. Pingshan and Dapeng are two management new districts subordinate to Longgang district.


Figure 4.1 (a) Shenzhen's location in China; (b) Relative locations of Shenzhen and Hong Kong (from Google Maps); (c) Shenzhen's administrative districts and management new districts. Nanshan, Futian, Luohu and Yantian are commonly known as *Guan Nei*, which are highly developed areas in terms of finance, education and tourism. The other six districts are usually known as *Guan Wai*, with manufacturing as its major industry.

commonly known as *Guan Nei*, which are highly developed areas in terms of finance, technology, education and tourism. The other six districts in Shenzhen are usually known as *Guan Wai*, with manufacturing as its major industry. According to the recent travel survey (Transport Commission of Shenzhen Municipality 2011), non-motorized trips accounted for a large percentage of the total trips in Shenzhen (walk: 50.0% and bicycle/moped: 6.2%). The city government considers bicycles as an effective transportation mode for people's daily travels, and plans to improve the corresponding facilities in the next few years. It is thus important to study where such facilities should be built to best accommodate people's travel needs.

This article uses an active tracking cellphone dataset[3] collected on a weekday in Shenzhen (03/23/2012). The dataset covers 5.8 million cellphones, with their locations reported approximately once every hour as (x, y) coordinates of the cellphone tower to which a cellphone is assigned. The dataset does not include location records for the 23:00 – 24:00 time window. Each cellphone, therefore, has 23 observations in a day. Table 4.1 shows an example of the data format of the cellphone dataset. The average nearest distance among the cellphone towers in this dataset is 0.19$km$.

Table 4.1 Example of an individual's cellphone location records.

| **User ID** | **Record ID** | **Time Window in which location was reported ($\boldsymbol{t}$)** | **Longitude of cellphone tower ($\boldsymbol{x}$)** | **Latitude of cellphone tower ($\boldsymbol{y}$)** |
|---|---|---|---|---|
| 932***** | 1 | [00:00 – 01:00) | 113.***** | 22.***** |
| 932***** | 2 | [01:00 – 02:00) | 113.***** | 22.***** |
| 932***** | 3 | [02:00 – 03:00) | 113.***** | 22.***** |
| **...** | **...** | **...** | 113.***** | 22.***** |
| 932***** | 23 | [22:00 – 23:00) | 113.***** | 22.***** |

[3] The mobile phone location dataset used in this study was acquired through research collaboration with Shenzhen Institutes of Advanced Technology, Chinese Academy of Sciences. The research was approved by the Institutional Review Board (IRB).

## 4.4 Methodology

This section first introduces how we generate important anchor points from individual cellphone trajectories. We then propose an anchor-point based trajectory segmentation method to partition the trajectories into trip chain segments. Then, we introduce how the potential demand of bicycle trips is generated from the derived trip chain segments. Finally we suggest facility locations of bike stations using a location-allocation model, and characterize the accessibility and temporal relationships between the incoming and outgoing trips at the bike stations.

### *4.4.1 Anchor point extraction and trajectory generalization*

As shown in Table 4.1, an individual's cellphone trajectory $V$ can be represented as:

$$T = \{\ P_1(x_1, y_1, t_1), P_2(x_2, y_2, t_2), \ldots, P_i(x_i, y_i, t_i)\ \} \tag{4.1}$$

where $P_i$ denotes the $i^{th}$ $(i = 1,2, \ldots, 23)$ cellphone location record; $x_i$ and $y_i$ denote the longitude and latitude of a cellphone tower, and $t_i$ represents an one-hour time window in which each location was recorded. Activity anchor point has been frequently used in the literature (e.g., Dijst 1999; Schönfelder and Axhausen 2003; Ahas et al. 2010) to denote a person's major activity locations such as home, workplace, favorite restaurants, etc. These activity anchor points serve as important activity origins/destinations of people's daily travels, and should be extracted to further estimate potential demand of bicycle trips. One challenge of using cellphone location data to determine an individual's activity anchor points is that an individual's cellphone location could switch among adjacent cellphone towers due to either cellphone load balancing (Csáji et al. 2013) or cellphone signal strength variation (Isaacman et al. 2012). Hence, it is necessary to consider these issues when estimating individual activity anchor points.

In this paper, we introduce *activity anchor point* as a set of cellphone towers that are geographically concentrated and where an individual spent a certain amount of time. A spatial clustering algorithm is performed to group the cellphone towers that are close to each other in space into clusters. In order to derive *activity anchor points* for $T$, we first extract all cellphone towers traversed by $T$, and calculate the frequency (i.e., number of time windows) of each cellphone tower that $T$ visited. We then select the most visited cellphone tower, and group all the cellphone towers that are located within $0.5km$ of the selected tower into a cluster. We then select the next most visited cellphone tower and perform the same grouping process. The process is repeated until all the cellphone towers in $T$ are processed. Finally, we calculate the number of cellphone location records (i.e., observations) assigned to each cluster. For each individual cellphone trajectory, any cluster with two or more cellphone location records are identified as an *activity anchor point*. The remaining clusters (i.e., isolated cellphone towers) are defined as *random cellphone towers*.

Note that we choose a constant distance threshold of $0.5km$ to derive activity anchor points from individual trajectories, and the reasons are as follows. First, although we are aware that cellphone tower densities could vary within a city, choosing a constant threshold enables us to consistently evaluate individuals' cellphone trajectories in a city. Second, as the average nearest distance among cellphone towers is $0.19km$ in Shenzhen, choosing $0.5km$ addresses the problem of signal switches among nearby cellphone towers, and uncovers individual movements among different activity clusters (i.e., inter-cluster movements).

Figure 4.2 gives an example of an individual's trajectory in a 3-dimensional space-time system proposed by Hägerstrand (1970). This individual's cellphone tower locations are grouped into four *clusters*, which include three *activity anchor points* (clusters A, B and C) and one *random cellphone tower* (cluster D). The red lines represent movements occurred within clusters (i.e., intra-cluster movements), and the green lines denote inter-cluster movements.

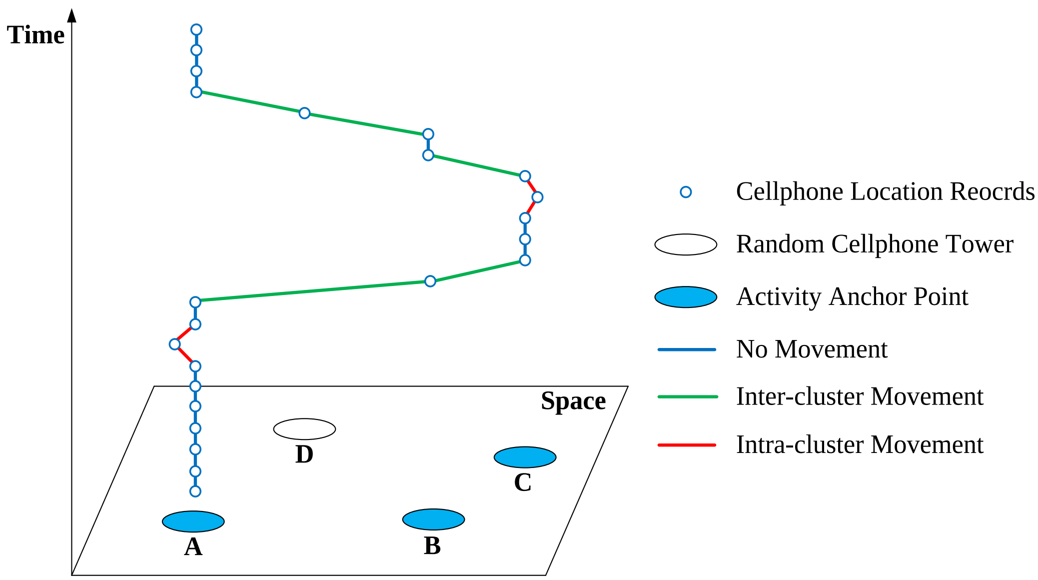



Figure 4.2 An individual's cellphone trajectory $T$ and concepts of activity anchor point and random cellphone tower.

Note that intra-cluster movements defined in our study could be caused by issues of cellphone signal switches, or refer to individual trips that were very short in distance. These intra-cluster movements are thus not considered as potential bicycle trips in the study. Hence, we merge cellphone towers in each cluster to derive a generalized cellphone trajectory. We choose the cellphone tower with the highest frequency in each cluster as the representative cellphone tower, and use these cellphone towers to generalize an individual's cellphone trajectory $T$. As illustrated in Figure 4.3, for a raw cellphone trajectory $T$, four representative cellphone towers (e.g., Tower A, B, C and D) that correspond to the four clusters are used to derive the generalized cellphone trajectory $T'$. The generalized cellphone trajectories in the dataset are used to derive individual trip chain segments, which are further used to estimate potential demand of bicycle trips.

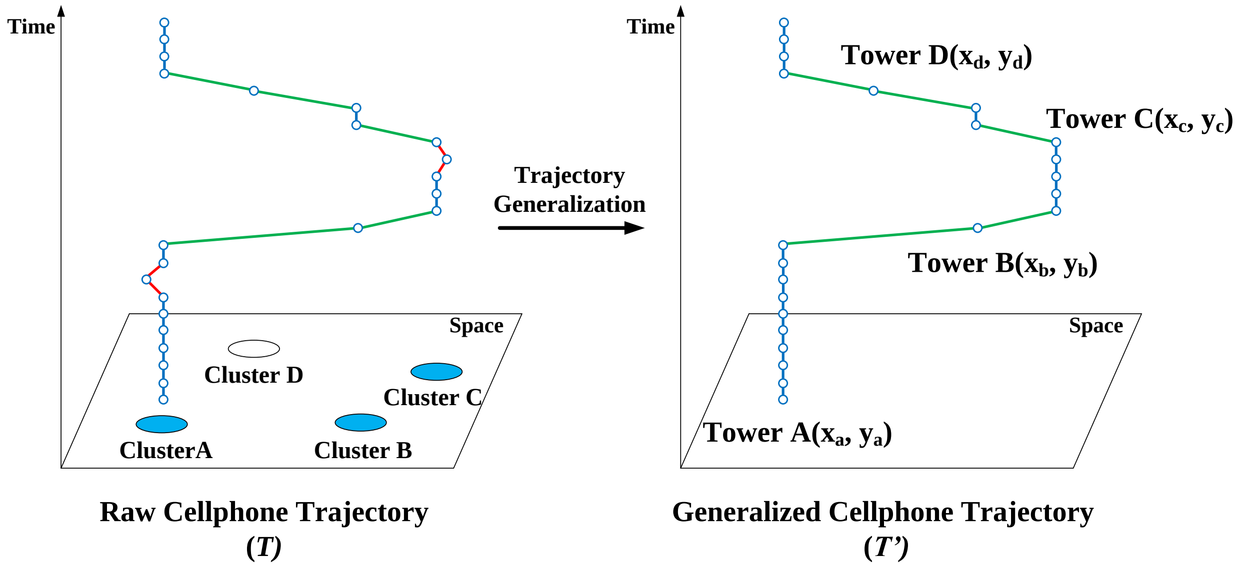



Figure 4.3 Derive generalized cellphone trajectory ($T'$) from an individual's raw cellphone trajectory ($T$) using representative cellphone tower in each cluster.

### *4.4.2 Trajectory segmentation based on trip chain analysis*

The trip-chaining behavior often describes an individual's daily travels that start and end at his/her major activity locations (e.g., home and workplace). The trip-chaining behavior reveals different types of human activity patterns and has great implications to transportation planning. In this study, we estimate two important activity anchor points for each individual, the *night-time anchor point* ($NTA$) and *day-time anchor point* ($DTA$), as approximate individual home location and workplace. These two anchor points serve as important activity origins/destinations when individuals plan their daily trips with single or multiple activity purposes. Hence, these two anchor points can be used to partition individual cellphone trajectories into meaningful segments for estimating the potential demand of bicycle trips.

According to Xu, Ling and Hu (2014), the normal hours of work and sleep for people in Shenzhen are 09:00 to 18:00 and 00:00 to 07:00, respectively. For each individual, the duration of stay at different cellphone towers during these two time periods are used to identify his/her *day-time anchor point* and *night-time anchor point*. By considering people's daily routines in most big cities in China, we adopt the approach proposed in (Long, Zhang and Cui 2012) to derive the two anchor points. In particular, we define an individual's *day-time anchor point* ($DTA$) as the activity anchor point with a minimum of 6 hours of stay at the location between 09:00 and 18:00. *Night-time anchor point* ($DTA$) is defined as the activity anchor point with a minimum of 4 hours of stay at the location between 00:00 and 07:00.

Based on this rule, we are able to estimate $NTA$ for 99% of the sampled population, and $DTA$ for 85% of the sampled population. As illustrated in Figure 4.4, 55% of the individuals have both $NTA$ and $DTA$ extracted, which correspond to different cellphone tower locations in the study area; 30% of the individuals have both $NTA$ and $DTA$ extracted that correspond to the same cellphone tower locations (i.e., $NTA$ and $DTA$ are co-located in space); 14% of the individual have only $NTA$ extracted; and 1% of the individuals have neither of the anchor points detected. In this study, individuals with neither of the two activity anchor points extracted (i.e., Type D in Figure 4.4) are removed from the remaining analysis.

We then partition the generalized cellphone trajectories into trip chain segments based on these two activity anchor points. For an individual's generalized cellphone trajectory $T'$, each trip chain segment after partition refers to a list of consecutive cellphone records, which originated and ended at either of the two important activity anchor points (i.e, $NTA$ and $DTA$). Table 4.2

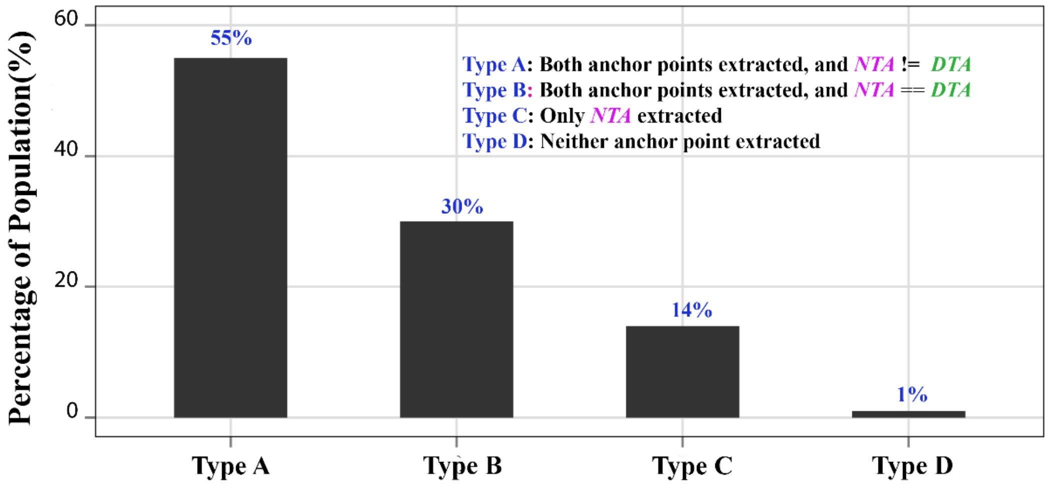



Figure 4.4 Percentage of population by four types of relationship between individual $NTA$ (night-time anchor point) and $DTA$ (day-time anchor point).

shows the four main types of trip chain segments derived in this study. $ND$ refers to the segments that started at individual $NTA$ and ended at individual $DTA$. The "$Passby$" locations refers to other cellphone towers (e.g., other activity anchor points or random cellphone towers) traversed by the corresponding trip chain segment. Similarly, $NN$ refers to the trip chain segments that both started and ended at $NTA$; $DN$ refers to the trip chain segments that started at $DTA$ and ended at $NTA$; $DD$ denotes the segments that both started and ended at $DTA$. Note that "$Passby$" locations do not always exist in $ND$ and $DN$ trip chain segments. For example, an individual could be located at $NTA$ during a certain one-hour time window and at $DTA$ during the next time window.

Note that two types of conditions are treated with special consideration during the trajectory segmentation. First, for individuals with $NTA$ and $DTA$ represented by the same cellphone tower (i.e., Type B shown in Figure 4.4), it is likely that these individuals stayed around one particular location (e.g., home) during the entire day. We thus consider all trip chain segments of these individuals as $NN$. Second, individuals did not always return to $NTA$ during the last one-hour time window (i.e., 22:00 – 23:00) in the study day. For each individual, if the last trip chain segment did not end at $NTA$, we add a virtual $NTA$ at the end of this segment by assuming that he/she would finally return to the night-time anchor point.

Table 4.2 Four types of trip chain segments derived from individual cellphone trajectories.

| **Type** | **Representation of Trip Chain Segment** | **Typical Activity Patterns** |
|---|---|---|
| $\boldsymbol{ND}$ | $NTA - Passby - DTA$ | Home – (Drop mail to FedEx) – Workplace |
| $\boldsymbol{NN}$ | $NTA - Passby - NTA$ | Home – (Shop at grocery store) – Home |
| $\boldsymbol{DN}$ | $DTA - Passby - NTA$ | Workplace – (Dine at restaurant) – Home |
| $\boldsymbol{DD}$ | $DTA - Passby - DTA$ | Workplace – (Meet with others at Starbucks ) – Workplace |

### *4.4.3 Generate potential demand*

This section introduces how we generate potential demand of bicycle trips from the four types of trip chain segments. According to the most recent travel survey conducted in Shenzhen (Transport Commission of Shenzhen Municipality 2011), the average trip distances for walking and bicycles are 1.6$km$ and 4.8$km$, respectively. However, it is pointed out in the survey that the average walking trip distance in Shenzhen is generally higher than that of other domestic and foreign cities (usually $1km$) due to the underdevelopment of cycling facilities and services. Hence, we consider $1km$ as the reasonable walking range, and use $1km$ and $5km$ as the spatial thresholds to filter the trip chain segments.

For all the trip chain segments extracted from an individual's generalized cellphone trajectory $T'$, we first calculate the *range* of each segment. The *range* of a segment is defined as the maximum distance (i.e., shortest path distance along road network) between all pairs of cellphone towers traversed by the segment. In our analysis, the trip chain segments with $range < 1km$, or $range > 5km$ are removed. We use this filtering strategy to exclude those trip chain segments that are either within reasonable walking range, or beyond normal travel distance for bicycles.

The reason of using *range* to filter each segment is that an individual might stop by other activity locations (i.e.., $Passby$ locations) during a trip chain segment defined in our study. If any of the two locations (i.e., cellphone towers) in a segment are beyond a distance threshold (e.g., $5km$), that means a certain part of the trip chain segment might not be well served by bicycles. For example, an individual needs to stop by two locations $P$ and $Q$ in a $NN$ segment (i.e., $NTA$ – $P$ – $Q$ – $NTA$). If the shortest path distance between $NTA$ and either of these two locations ($P$ and $Q$) is greater than $5km$, the individual is not likely to use bicycle for this trip chain segment. We apply this filtering strategy to the four types of trip chain segments in our study.

As individual cellphone trajectories were recorded at the cellphone tower level, the potential demand is thus aggregated by individual cellphone towers. In our analysis, two basic types of demand, $inflow_p$ and $outflow_p$, are extracted at each cellphone tower $p$ during different time periods of the study day:

$$inflow_p = (I_1^p, I_2^p, I_3^p, \dots, I_{22}^p) \tag{4.2}$$

$$outflow_p = (O_1^p, O_2^p, O_3^p, \dots, O_{22}^p) \tag{4.3}$$

$$total\_inflow_p = \sum_{i=1}^{22} I_i^p \tag{4.4}$$

$$total\_outflow_p = \sum_{i=1}^{22} O_i^p \tag{4.5}$$

As shown in equation 4.2 and 4.3, $I_i^p$ and $O_i^p$ refer to the volume of incoming and outgoing trips at cellphone tower $p$ during a particular time interval $i$, respectively. For example, time interval 1 denotes the time interval between time windows $t_1$ (00:00 – 01:00) and $t_2$ (01:00 – 02:00). As illustrated in Table 4.1, each cellphone trajectory covers 23 time windows (i.e., observations) in the study day. Hence, the $inflow_p$ and $outflow_p$ each has 22 observations.

Note that $total\_inflow_p$ and $total\_outflow_p$ refer to the total amount of incoming and outgoing trips at cellphone tower $p$ for the entire day, respectively. The two measures will be used later as the input for a location-allocation model to determine the facility locations of bike sharing stations.

We next introduce how $inflow_p$ and $outflow_p$ are extracted from all trip chain segments with $range$ between $1km$ and $5km$. Note that a trip chain segment $TS$ can be represented as a series of cellphone tower locations:

$$TS = \{\, P_1(x_1, y_1, t_1), P_2(x_2, y_2, t_2), \dots, P_i(x_i, y_i, t_i)\} \tag{4.6}$$

where $P_i$ denotes individual cellphone tower, $x_i$ and $y_i$ denote the longitude and latitude of $P_i$, and $t_i$ represents the one-hour time window when the cellphone location was recorded. By comparing each pair of consecutive cellphone towers ($P_i$ and $P_{i+1}$) in $TS$, we assign a unit of demand to $outflow_{Pi}$ at the corresponding time interval (i.e., the time interval associated to $t_i$ and $t_{1+1}$), and a unit of demand to $inflow_{Pi+1}$ if:

$$\begin{cases} x_i \neq x_{i+1} \\ y_i \neq y_{i+1} \end{cases} \tag{4.7}$$

We repeat this calculation until all trip chain segments ($ND$, $NN$, $DN$ and $DD$) in the dataset have been processed.

### *4.4.4 Suggest facility locations of bike stations*

This study applies a location-allocation model to suggest facility locations of bike sharing stations. In particular, the location-allocation module included in the Network Analyst toolbox of ArcGIS 10.1 is used. The location-allocation module in ArcGIS 10.1 offers different problem types (e.g., minimize impedance, maximum coverage and minimize facilities) to answer specific kinds of questions. This study uses the *maximum coverage* module in ArcGIS 10.1 to suggest locations of bike sharing stations. The objective of this module is to locate a fixed number of facilities (bike stations in this study) such that the total demand within a specified impedance cutoff (i.e., service radius) of the facilities is maximized.

When configuring the *maximum coverage* module, the individual cellphone towers in the actively tracked cellphone location dataset (5928 in total) are used as both *demand points* and the *candidate locations* of the facilities (Appendix 2 and Appendix 3). The *weight* at each demand point $p$ (i.e., cellphone tower $p$) is calculated as the sum of $total\_inflow_p$ and $total\_outflow_p$ (Appendix 4) since they correspond to the number of drop-off and pick-up activities of potential bicycle trips, respectively. These two types of activities are both considered as travel demand when planning bike sharing stations in a city (García-Palomares, Gutiérrez and Latorre 2012).

The *impedance cutoff* is chosen as 500 meters (road network distance) to approximate the service radius of the facilities (Appendix 1). For the *number of facilities* ($N$) to be located, we create four different scenarios (i.e., $N = 300$; $N = 600$; $N = 900$ and $N = 1200$) and compare the potential outcome (e.g., percentages of demand that can be covered) among different solutions.

Once the locations of the facilities have been determined in each of the four scenarios, the location-allocation module will allocate the demand points to the facilities. In particular, any demand point which falls outside of all facilities' impedance cutoffs is not allocated. A demand point which is inside the impedance cutoff of one facility is allocated to that facility, while a demand point which falls within the impedance cutoff of two or more facilities is allocated to its nearest facility.

### 4.4.5 Characterization of bike stations

In this study, two measures are introduced to assess the bike stations once their locations are derived from the location-allocation model. First, we introduce an accessibility measure to evaluate how well the stations could serve bicycle users to reach other potential activity destinations. We also investigate the dynamic relationships between the incoming and outgoing trips at the bike stations over time. The dynamic relationships serve as useful information for the distribution and redistribution of bicycles among the bike stations during different time periods of a day.

In order to measure the two important characteristics of the bike stations, we first retrieve the demand points that are allocated to each bike station, and then calculate the total amount of demand allocated to each bike station after the location-allocation model is performed.

Specifically, for each bike station $q$, we introduce $inflow\_C_q$ and $outflow\_C_q$ to represent the amount of incoming and outgoing trips that are allocated to the station:

$$inflow_q = (J_1^q, J_2^q, J_3^q, \dots, J_{22}^q) \tag{4.8}$$

$$outflow_q = (K_1^q, K_2^q, K_3^q, \dots, K_{22}^q) \tag{4.9}$$

For each bike station $q$, $J_i^q$ and $K_i^q$ refer to the number of incoming and outgoing trips at the demand points that are allocated to $q$ during time interval $i$ (e.g., 1, 2, 3, … 22), respectively:

$$J_i^q = \sum_{m=1}^{n} O_i^m * C_{qm} \tag{4.10}$$

$$K_i^q = \sum_{m=1}^{n} I_i^m * C_{qm} \tag{4.11}$$

Here $n$ denotes the total number of demand points (i.e., cellphone tower locations) in the study area. $C_{qm}$ takes the value of 1 if demand point $m$ is allocated to the bike station $q$, and 0 otherwise. Note that:

$$total\_inflow\_C_q = \sum_{i=1}^{22} J_i^q \tag{4.12}$$

$$total\_outflow\_C_q = \sum_{i=1}^{22} K_i^q \tag{4.13}$$

By doing so, we are able to aggregate the incoming and outgoing trips from the demand points to the bike stations and maintain the temporal characteristics of the demand.

The concept of accessibility has been widely used in transportation studies to describe how well locations could reach other potential activity destinations (Hansen 1959). In order to represent the accessibility of each bike station, we adopt a gravity-based measure that has been used in previous studies to quantify bicycle accessibility (García-Palomares, Gutiérrez and Latorre 2012; Iacono, Krizek and El-Geneidy 2012). For each bike station $q$, the accessibility $A_q$ is calculated as follows:

$$A_q = \sum_{k=1}^{n} \frac{total\_inflow\_C_k * M_{qk}}{(D_{qk})^{\alpha}} \tag{4.14}$$

Here $n$ denotes the total number of bike stations (e.g., 300, 600, 900 and 1200) in our analysis. $M_{qk}$ takes the value of 1 if the road network distance between station $q$ and station $k$ is less than $5km$, and is 0 otherwise. $D_{qk}$ is the road network distance between station $q$ and station $k$, and $\alpha$ takes the value of 2 (which is the default value of the gravity based measure) to reflect the distance decay effect. Note that we use $total\_inflow\_C_k$ (i.e., number of incoming trips) to approximate the total amount of opportunities (i.e., activities) at station $k$.

For each station $q$, we introduce $netflow_q$ to represent the dynamic relationships between incoming and outgoing trips:

$$netflow_q = (N_1^q, N_2^q, N_3^q, \dots, N_{22}^q) \quad (4.15)$$

For each time particular time interval $i$, $N_i^q$ is calculated as the net volume of trips normalized by the total trips:

$$N_i^q = \frac{K_i^q - J_i^q}{K_i^q + J_i^q} \quad (4.16)$$

The value of $N_i^q$ ranges from -1.0 to 1.0. A positive $N_i^q$ indicates that station $q$ served as a *trip producer* at time interval $i$, while a negative value indicates that station $q$ served as a *trip attractor* during that time interval. The temporal characteristics of $netflow_q$ describes the asymmetry of human travel patterns at different times of a day. In order to assess the temporal characteristics of $netflow_q$, this study uses a k-means algorithm to group the bike stations into clusters. The clustering results are used to better understand the varying temporal characteristics of $netflow_q$ at the bike stations and their corresponding geographic distributions.

## 4.5 Analysis Results

### *4.5.1 General statistics*

By analyzing the generalized cellphone trajectories of 5.8 million individuals in the dataset, we were able to derive four types of trip chain segments. As shown in Table 4.3, we extracted 1,636,494 $ND$ segments (24.3% of total) and 1,480,342 $DN$ segments (22.0% of total). The percentages of $ND$ and $DN$ segments are close to each other, which reflects the regularity of human travels between their day-time and night-time anchor points during the study day. The number of $NN$ segments is 3,159,753 (47.0% of total), which indicates that trips around individual night-time anchor point serve as a major component in people's daily travels in Shenzhen. We also identified 449,652 $DD$ segments, which account for only 6.7% of total segments.

Table 4.3 Number and percentage of extracted trip chain segment by type.

| **Type of Trip Chain Segment** | **Amount** | **Percentage of Grand Total** |
|---|---|---|
| $ND$ | 1,636,494 | 24.3% |
| $NN$ | 3,159,753 | 47.0% |
| $DN$ | 1,480,342 | 22.0% |
| $DD$ | 449,652 | 6.7% |

We next examine the duration (i.e., number of time windows covered by a segment) of the trip chain segments by type. As shown in Figure 4.5a, most of the $ND$ trip segments have a duration of two (about 80% of total) or three time windows (about 15% of total). Similar distribution pattern is observed for $DN$ segments (Figure 4.5c). It is very likely that individuals tended not to visit or stay at other activity destinations for a long time during their travels between night-time and day-time anchor points. As illustrated in Figure 4.5b and 4.5d, a considerable proportion of $NN$ and $DD$ segments have a duration of three or four time windows. However, the long tail in Figure 4.5b indicates that certain individuals stayed at particular activity destinations for a relatively long time during $NN$ segments.

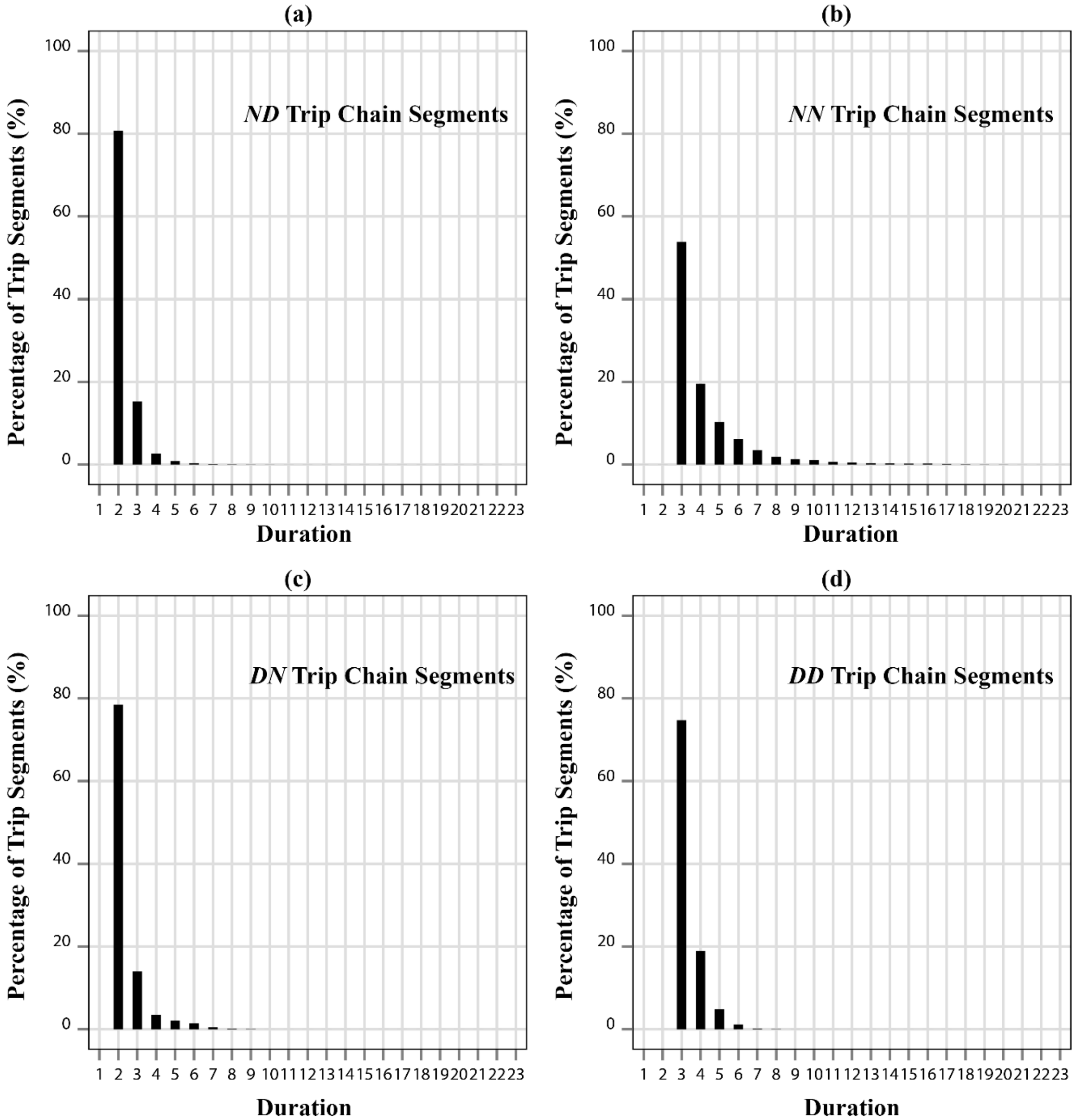



Figure 4.5 Distribution patterns of the duration of: (a) $ND$ trip chain segments; (b) $NN$ trip chain segments; (3) $DN$ trip chain segments; (4) $DD$ trip chain segments.

### *4.5.2 Temporal distributions of trip chain segments*

Next we explore the temporal variations of the four types of trip chain segments. Figure 4.6a illustrates the temporal distribution of $ND$ segments (horizontal axis represents the starting time window of the corresponding segments, and vertical axis denotes the volume of segments). We can see that the majority of segments occurred during morning rush hours. This is because $ND$ segments mainly describe people's commuting activities during these time periods (i.e., time windows 7, 8 and 9). We also notice that there was a local peak at time window 13. This can be potentially explained by the activity patterns of particular people in Shenzhen, who usually went back home from their workplaces in order to have a short rest during noon time. Similar temporal patterns are observed for $DN$ segments (Figure 4.6c). The number of $DN$ segments reached its peak around afternoon rush hours but decayed slowly during the night time. The identified patterns can be potentially explained by two reasons: (1) people chose different time to get off work to avoid traffic congestions; (2) some people might need to work overtime, and left their workplaces late in the evening. Even though $ND$ and $DN$ segments refer to travels that were relatively short in distance, we notice that the temporal patterns still reflect the rhythms of urban dynamics.

As illustrated in Figure 4.6b, the volume of $NN$ segments were relatively consistent over time in the day. The result indicates that activities around people's $NTA$ played an important role in manifesting the urban dynamics in Shenzhen. As shown in Figure 4.6d, the volume of $DD$ segments mainly concentrated during normal work hours with its peak around time interval 12, which indicates higher intensity of human activities (e.g., dining) around individual $DTA$ during noon time.

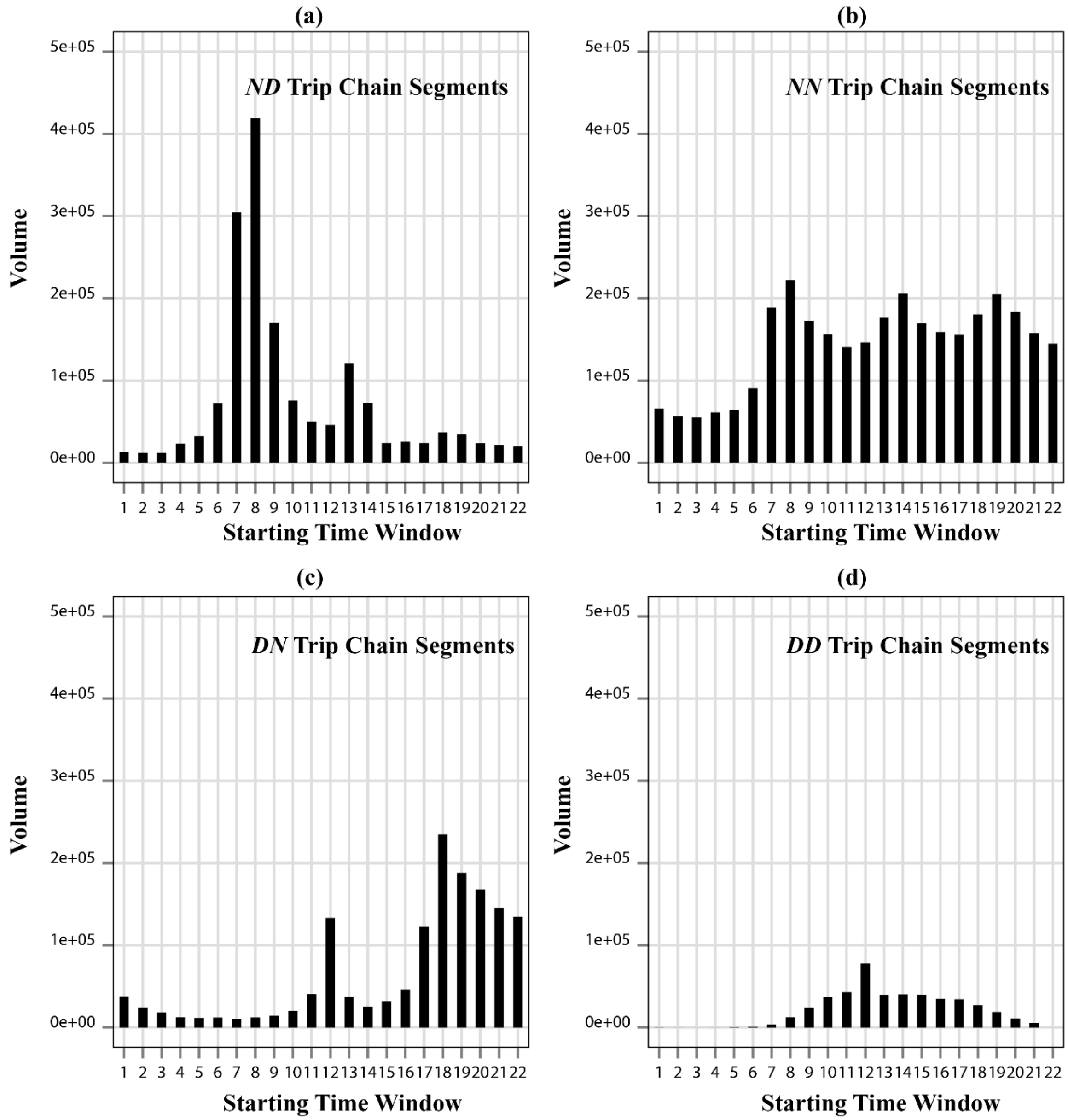



Figure 4.6 Temporal distributions of trip chain segments by type: (a) $ND$ trip chain segments; (b) $NN$ trip chain segments; (3) $DN$ trip chain segments; (4) $DD$ trip chain segments.

*4.5.3 Spatial and temporal dynamics of $outflow_p$ and $inflow_p$*

Understanding the spatial and temporal dynamics of potential demand serves as critical information for the planning and operation of bike sharing stations. As $outflow_p$ and $inflow_p$ were generated at the cellphone tower level, we use kernel density maps to illustrate the geographic distributions of the demand at different times of the day. As a bike sharing station only serve the demand points (i.e., cellphone towers) that are nearby, a small search radius is needed to fit a density surface to properly reflect the geographic patterns of potential demand. By comparing the kernel density maps using two search radii ($0.5km$ and $1km$), we finally chose $1km$ as the search radius which produced better visualization effect.

In this section, several key time intervals are chosen to illustrate the geographic distributions of $outflow_p$ and $inflow_p$. For example, Figure 4.7a shows the density patterns of $outflow_p$ at time interval 8 (i.e., 07:00 – 08:00 to 08:00 – 09:00). Areas with a high density of demand mainly locate at south Futian, southwest Bao'an, southwest Nanshan, and central Longhua. These areas generated large number of potential bicycle trips in the early morning. By further overlaying the density map with the land use patterns, we find that these areas mainly refer to residential neighborhoods in Shenzhen. For example, areas $A$, $B$, $C$, $F$ and $G$ refer to places with many residential apartments. Area $D$ and $E$ cover several "urban villages" (e.g., Shangsha Village and Huanggang Village) in Shenzhen. These "urban villages" (Wei and Yan 2005) usually refer to densely populated areas with large migrant population.

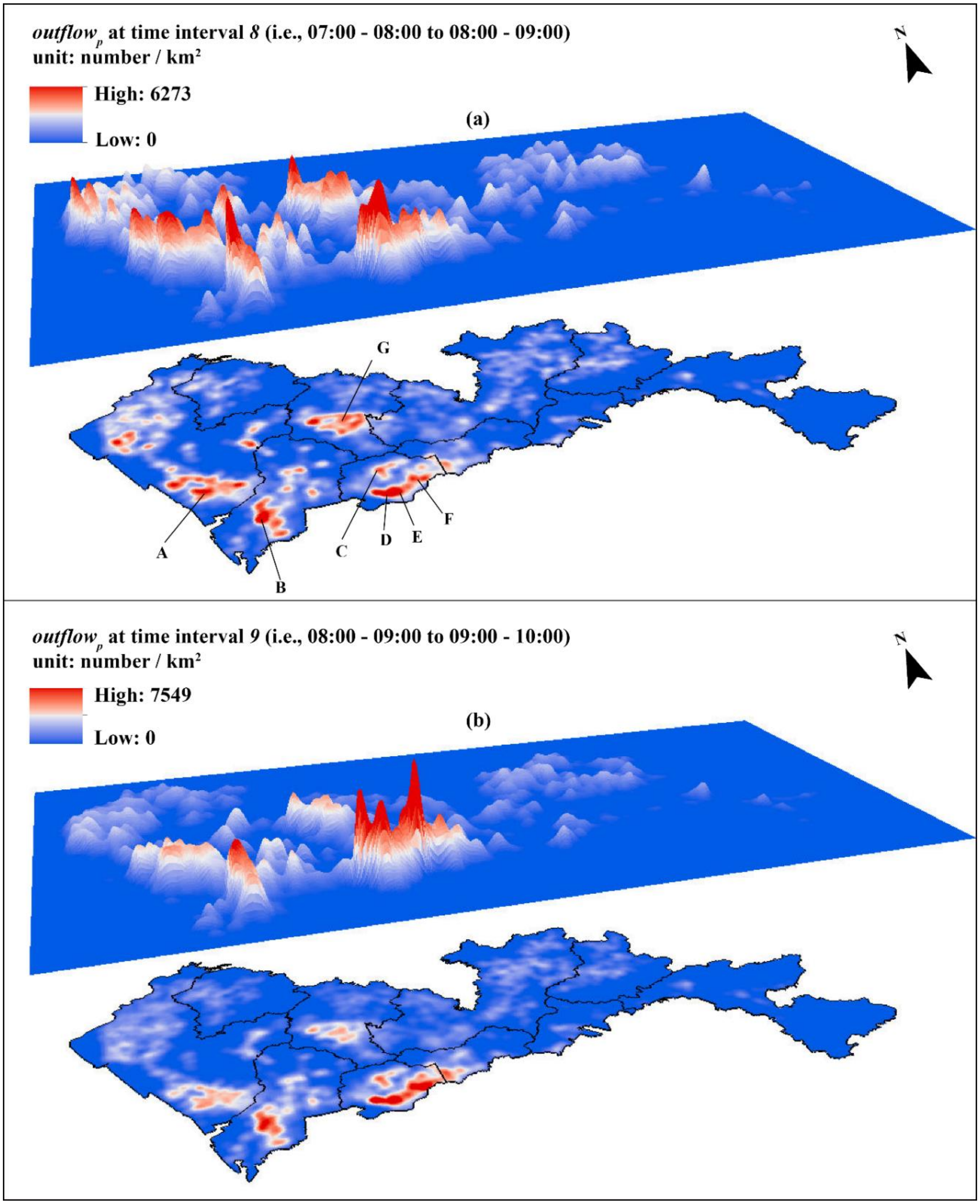



Figure 4.7 (a) Spatial distribution patterns of $outflow_p$ during time interval 8; (b) Spatial distribution patterns of $outflow_p$ during time interval 9. The upper maps show the 3D view of the density patterns.

Figure 4.7b shows the density patterns of $outflow_p$ at time interval 9 (i.e., 08:00 – 09:00 to 09:00 –10:00). We notice that certain areas in south Nanshan and south Futian still generated large number of trips, while the intensities of $outflow_p$ became lower in the northern part of Shenzhen as compared to the previous time interval. As described previously, the northern districts (i.e., *Guan Wai*) in Shenzhen are mainly industrial-oriented areas with a large number of migrant workers, while the districts in the south (i.e., *Guan Nei*) offer more employment opportunities related to education, technology and commerce. The identified patterns could be potentially explained by the differences of people's work schedules between *Guan Nei* and *Guan Wai* with varying economic and employment structures.

Figure 4.8a illustrates the density patterns of $inflow_p$ at time interval 8 (i.e., 07:00 – 08:00 to 08:00 – 09:00). Several areas with a high density of demand are highlighted in the map. We notice that certain industrial parks (e.g., Foxconn Technology Park in Longhua and Yintian Industrial Park in Bao'an) in the northern districts attracted large number of trips in the early morning. In southern Shenzhen, however, the areas with a high density of demand mainly refer to commercial districts and business centers. Figure 4.8b shows the density patterns of $inflow_p$ at time interval 9 (i.e., 08:00 – 09:00 to 09:00 – 10:00). We notice that the industrial parks in the north attracted much fewer trips during time interval 9 as compared to time interval 8, which is consistent with our findings in Figure 4.7. However, the commercial areas and business centers in Futian and Luohu districts continued to attract large number of trips. The area with the highest density refers to Huaqiang North, which is the largest commercial center in Shenzhen, and is famous for its business of computer hardware and electronic products.

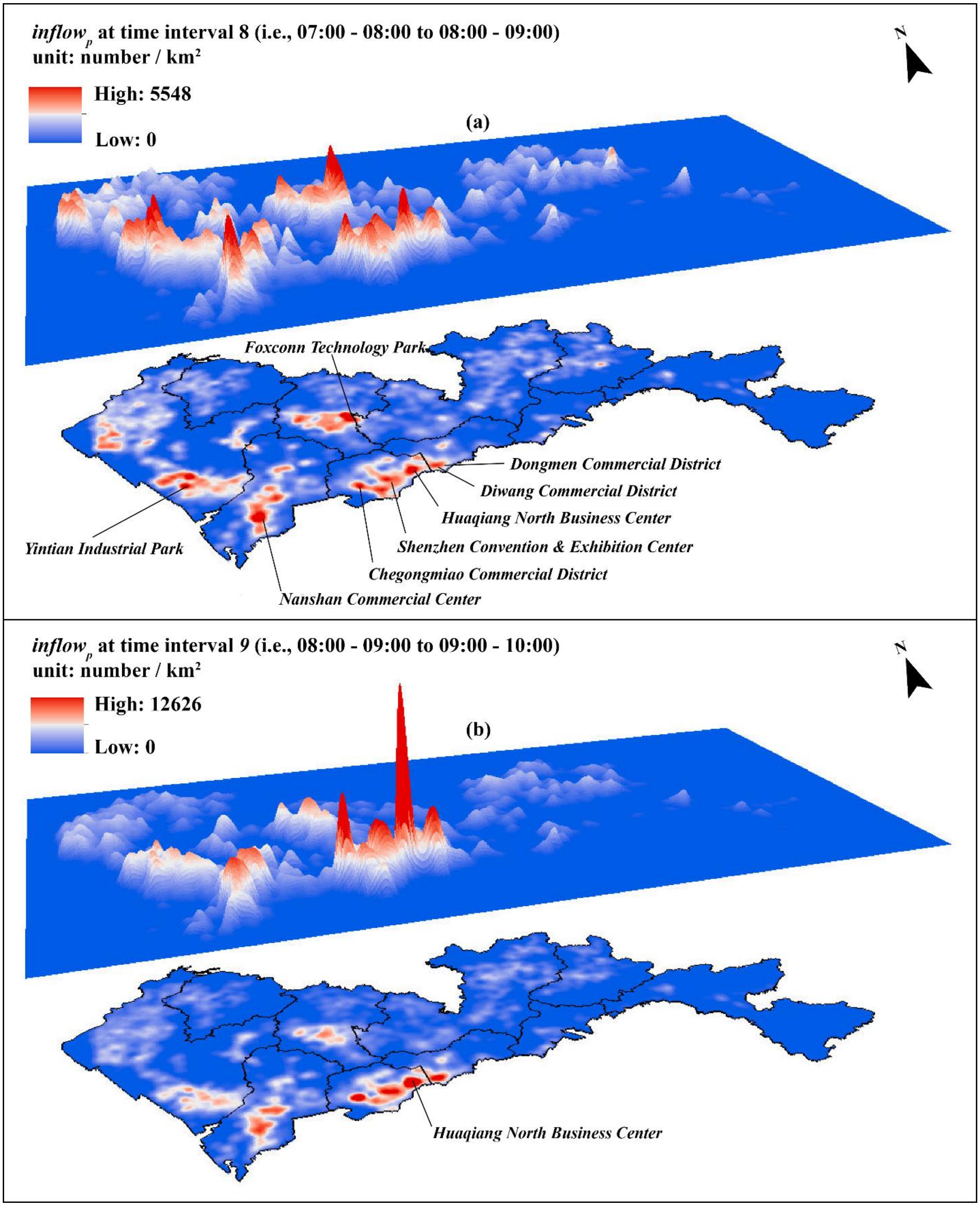



Figure 4.8 (a) Spatial distribution patterns of $inflow_p$ during time interval 8; (b) Spatial distribution patterns of $inflow_p$ during time interval 9.

Figure 4.9 illustrates the geographic patterns of $outflow_p$ and $inflow_p$ at time interval 18 (i.e., 17:00 – 18:00 to 18:00 – 19:00), respectively. It can be perceived that the areas which attracted large number of trips during time interval 8 and 9 (Figure 4.8) also generated a significant amount of trips during time interval 18 (Figure 4.9a). Similarly, the areas that generated large number of trips during morning rush hours (Figure 4.7) also attracted many trips in the late afternoon (Figure 4.9b). The analysis results reflect the urban rhythms of human travel patterns in Shenzhen.

We next examine the density patterns of $outflow_p$ and $inflow_p$ at time interval 15 (i.e., 14:00 – 15:00 to 15:00 – 16:00). As morning and afternoon rush hours refer to the time periods when a large number of $ND$ and $DN$ segments occurred, we choose this particular time interval to better understand the dynamics of travel demand related to other types of trip chain segments (e.g., $NN$ trip segments). By comparing Figure 4.10a and 4.10b, we notice that the density patterns of $outflow_p$ and $inflow_p$ are very similar to each other at time interval 15. It can be perceived that the areas which generated more trips tended to also attract more trips at the same time. As shown in Figure 4.6, a considerable proportion of $outflow_p$ and $inflow_p$ at time interval 15 were extracted from $NN$ trip chain segments. By overlaying the density patterns with Google Maps, we find that many areas with a high density of $outflow_p$ (and $inflow_p$) are associated with recreational and shopping activities. For example, we find many parks (e.g., Longhua Park, Xixiang Park and Tiezaishan Park) with a high density of travel demand at time interval 15. These parks are free and open to the public, and offer various sports and recreational facilities. The Nanshan Cultural & Sports Center, funded by the local government, has several art schools, amateur sports schools, culture centers and theaters which provide different types of recreational activities. The Dongmen commercial area located in Luohu district integrates commerce, tourism,

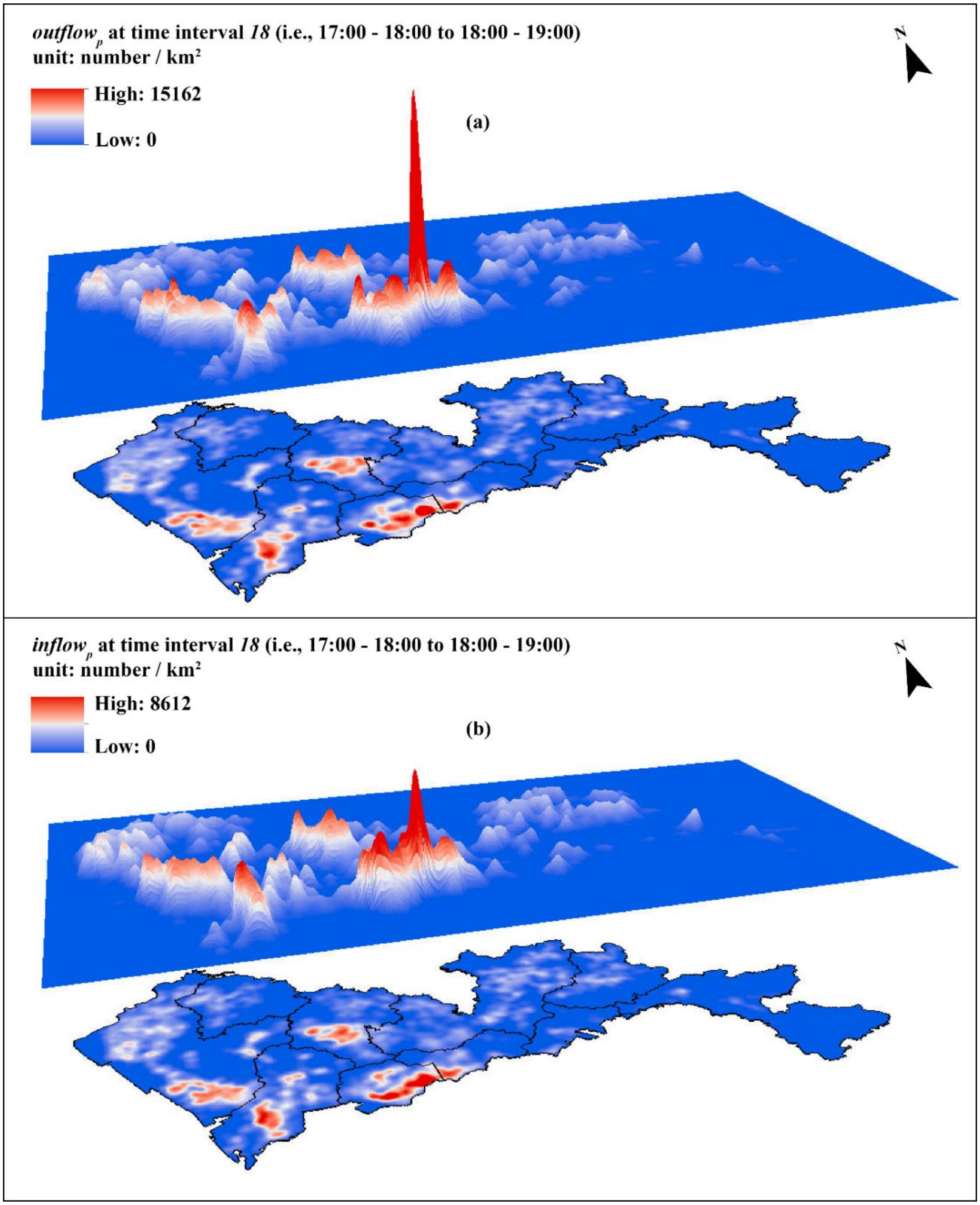



Figure 4.9 (a) Spatial distribution patterns of $outflow_p$ during time interval 18; (b) Spatial distribution patterns of $inflow_p$ during time interval 18.

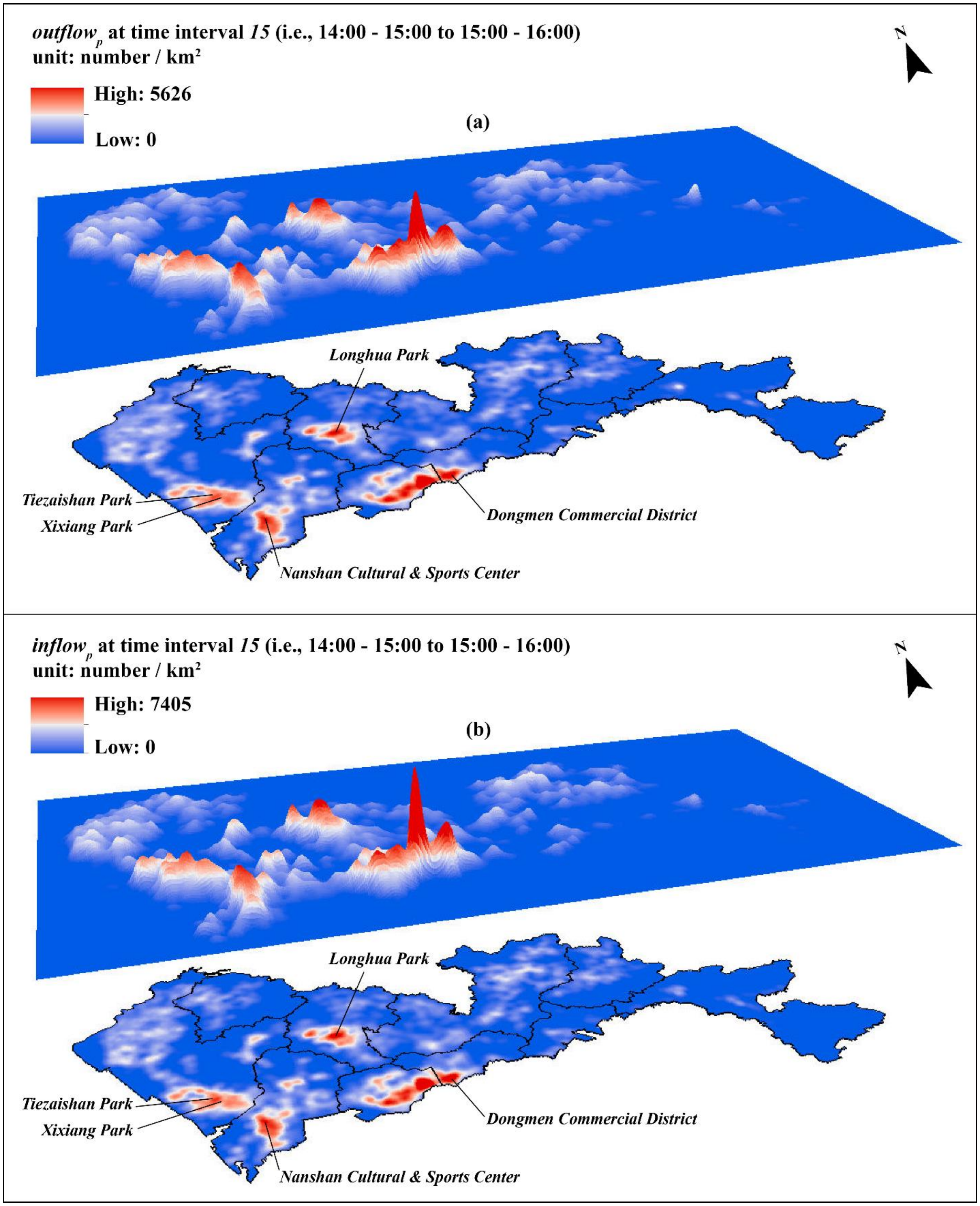



Figure 4.10 (a) Spatial distribution patterns of $outflow_p$ at time interval 15; (b) Spatial distribution patterns of $inflow_p$ at time interval 15.

shopping and recreation as its core functions. It is very likely that $NN$ trip chain segments are strongly tied to people's leisure activities.

### *4.5.4 Facility locations for bike stations*

In this study, 5928 unique cellphone towers in the dataset are used as both demand points and candidate facility locations. As described previously in section *4.4.4*, the total demand (i.e., *weight*) at each cellphone tower $p$ is calculated as the sum of $total\_outflow_p$ and $total\_inflow_p$. Figure 4.11 illustrates the geographic distribution of the cellphone towers and the density of total demand (using $1km$ as the search radius). Areas with high density of total demand mainly locate in central Longhua, southwest Bao'an, south Nanshan and Futian. These areas correspond to places with high population densities in Shenzhen.

Figure 4.12 shows the locations of the bike stations derived from the location-allocation model. Four different scenarios (300, 600, 900 and 1200 facility locations) are compared in our analysis. As illustrated in Figure 4.12a, when the number of facilities ($N$) is set to 300, the majority of bike stations are located around the areas with a high density of total demand (e.g., central Longhua, southwest Bao'an, southwest Nanshan, southwest Luohu, and Futian). As the number increases to 600 (Figure 4.12b), we find that most of the stations still tend to be located around those areas with high demand, and only a limited number of new stations are assigned to the northern part of Shenzhen. As we continue to change $N$ to 900 and 1200 (Figure 4.12c and 4.12d), we notice that the bike stations gradually cover certain areas in the northern part of Shenzhen (e.g., Guangming, Longgang and Pingshan Districts).

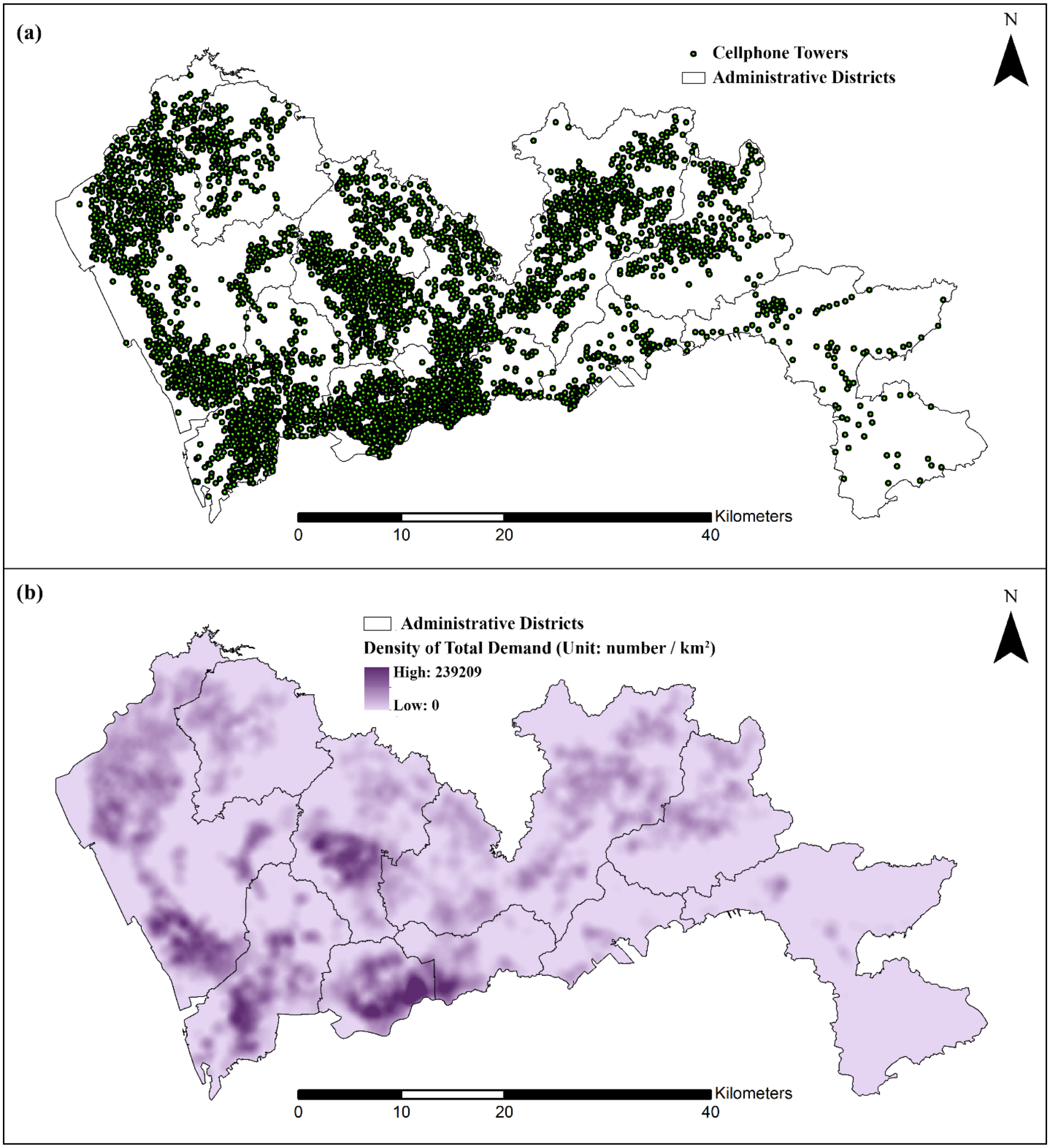



Figure 4.11 (a) Spatial distribution of cellphone towers (5928 in total); (b) Density of total demand (unit: $number\ /\ km^2$).

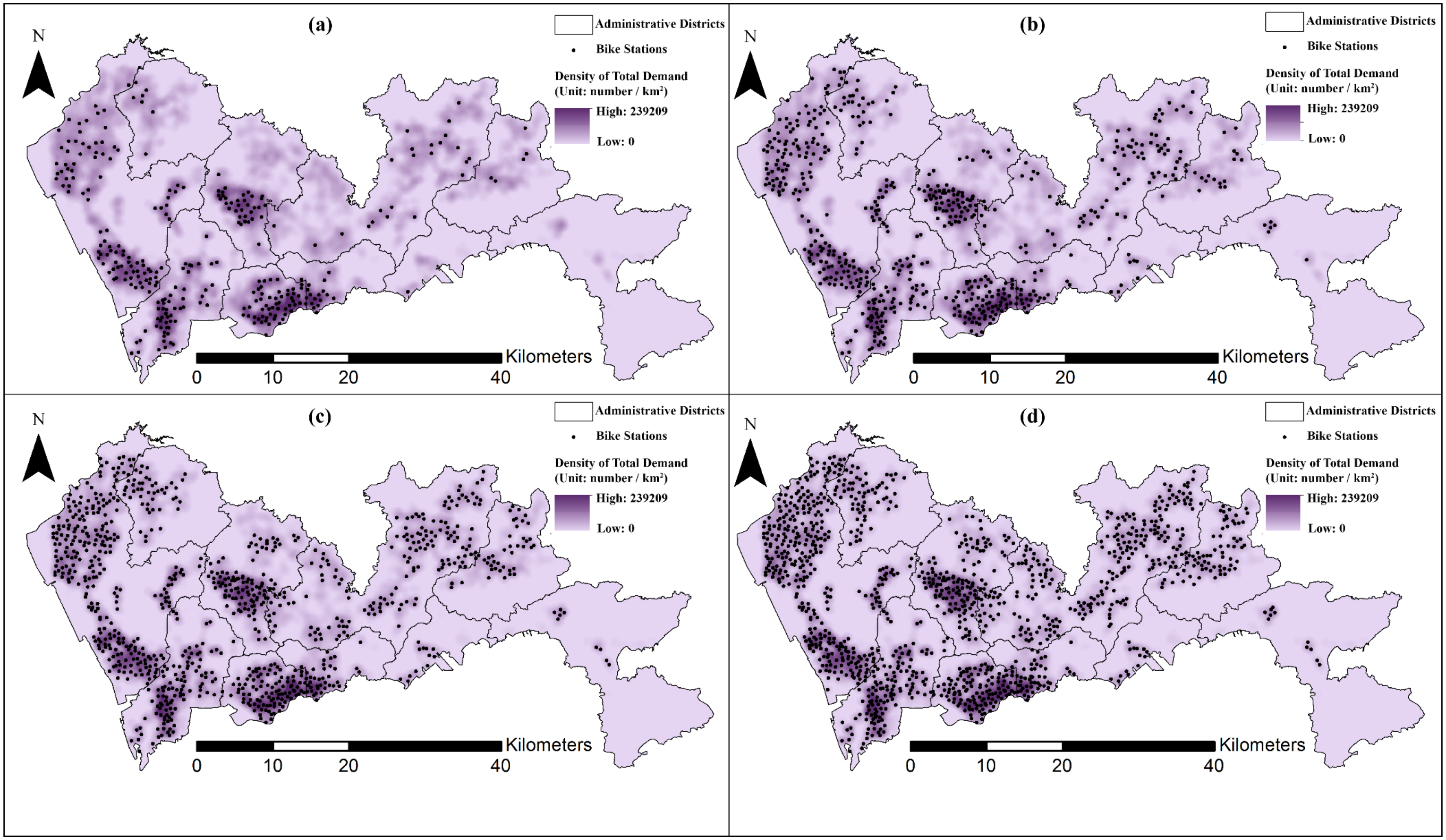



Figure 4.12 Facility locations of bike stations derived from location-allocation model: (a) 300 facilities; (b) 600 facilities; (c) 900 facilities and (d) 1200 facilities.

Table 4.4 summarizes the amount and percentage of total demand that can be covered by the facility locations in the four different scenarios. When $N$ equals 300, the percentage of total demand that can be covered is 40.2%. The solution with 300 bike stations covers a considerable proportion of total demand in Shenzhen because the majority of stations are located to serve the areas with a very high density of demand (Figure 4.12a). As $N$ increases from 300 to 1200, the percentage of total demand covered gradually increases from 40.2% to 84.6%. However, the increment percentage decreases from 20.2% ($N$ = 600) to 10.1% ($N$ = 1200), which shows a decay of new demand that can be covered as equal increment of new bike stations are added.

Table 4.4 Amount and percentage of total demand covered by the bike stations.

| Number of stations ($N$) | Amount Covered | Percentage of Amount Covered | Increment | Increment Percentage |
|---|---|---|---|---|
| 300 | 9,888,085 | 40.2% | – | – |
| 600 | 14,860,322 | 60.4% | 4,972,237 | 20.2% |
| 900 | 18,325,887 | 74.5% | 3,465,565 | 14.1% |
| 1200 | 20,829,777 | 84.6% | 2,503,890 | 10.1% |

### *4.5.5 Accessibility of the bike stations*

Figure 4.13 shows the accessibility of the bike stations in the four different scenarios. When $N$ = 300, bike stations with high accessibility are mainly located in the core areas (e.g., central Longhua, southwest Bao'an, southwest Nanshan, southwest Luohu, and the Futian) where the density of total demand is high (as shown in Figure 4.11b). As $N$ increases to 600, there is an increase in the overall accessibility of bike stations in these core areas. However, the majority of bike stations in the northern part of Shenzhen still experience low accessibility.

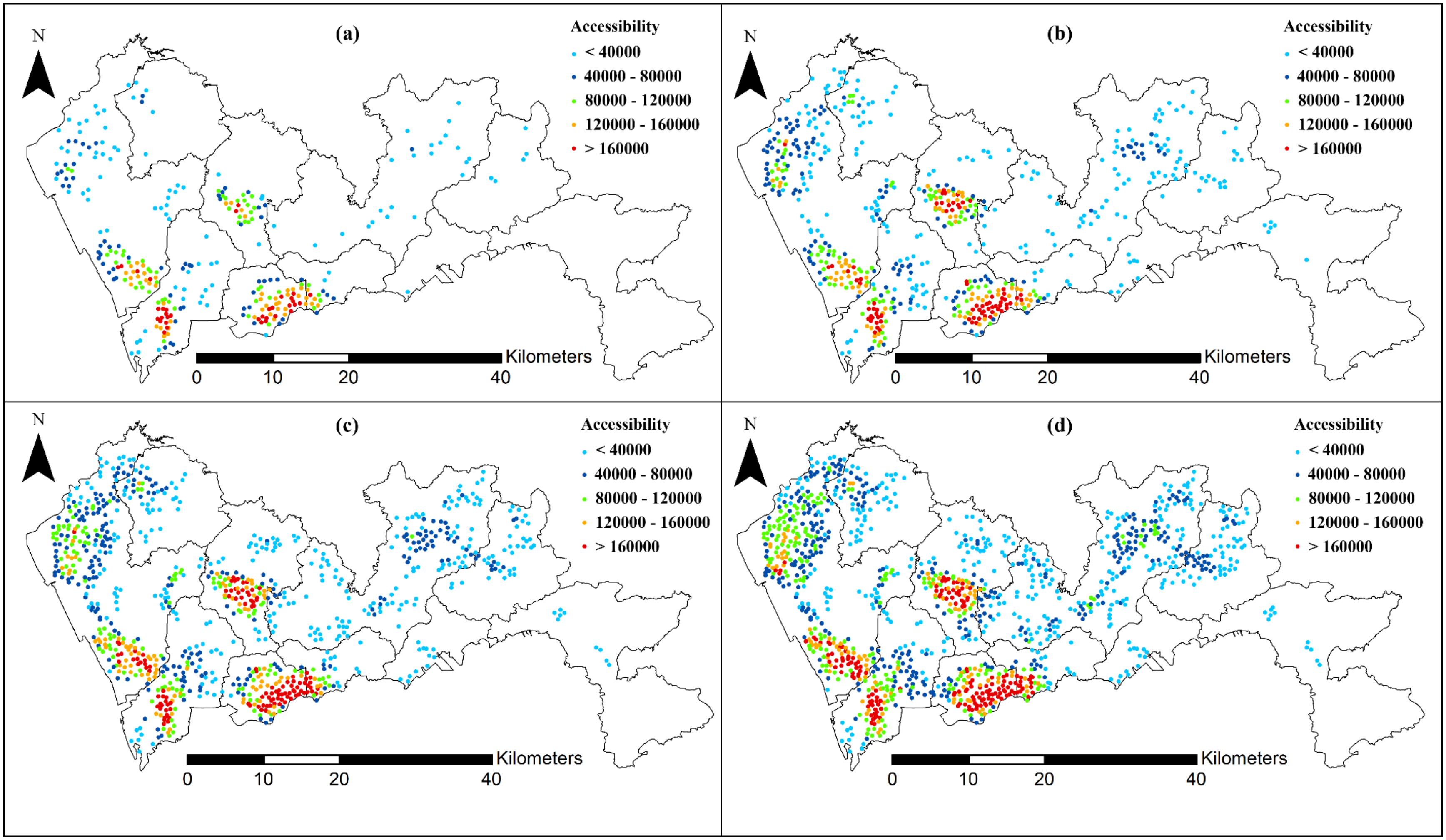



Figure 4.13 Accessibility of the bike stations: (a) 300 facilities; (b) 600 facilities; (c) 900 facilities and (d) 1200 facilities.

There is an increase of the accessibility of bike stations in the northern part of Shenzhen as $N$ changes from 600 to 900. As $N$ changes to 1200, we observe a slight increase of accessibility of the stations in certain areas but the trend is not obvious.

Figure 4.14 shows the average accessibility of bike stations by administrative districts in Shenzhen (Dapeng and Yantian are not included in this particular analysis due to low number of bike stations in the four scenarios). In general, bike stations in the southern districts (e.g., Futian, Luohu and Nanshan) have a higher average accessibility than the ones in the northern districts (e.g., Guangming, Longgang and Guangming). As $N$ changes from 300 to 1200, we observe an overall increase of average accessibility of bike stations in most of the districts. However, we notice that the average accessibility tends to decrease or remains stable for districts such as Futian, Nanshan and Longhua as $N$ changes from 900 to 1200. This is because when $N$ becomes very large, the new bike stations added to these districts tend to be located in peripheral areas where the density of demand is relatively low. On one hand, there are fewer potential activity destinations (i.e., opportunities) around these bike stations, which cause them to have relatively low level of accessibility. On the other hand, these newly added bike stations do not significantly improve the accessibility of nearby bike stations due to their low level of available opportunities (i.e., $total\_inflow\_C_q$ in our analysis). The analysis result in Figure 4.14 indicates that in districts where the potential demand concentrated in particular areas, adding more bike stations would result in a greater improvement (of average accessibility) at the beginning but diminishing returns when $N$ becomes larger. However, for districts where potential demand is less heterogeneous over space, adding more bike stations would enhance the overall accessibility of the stations smoothly.

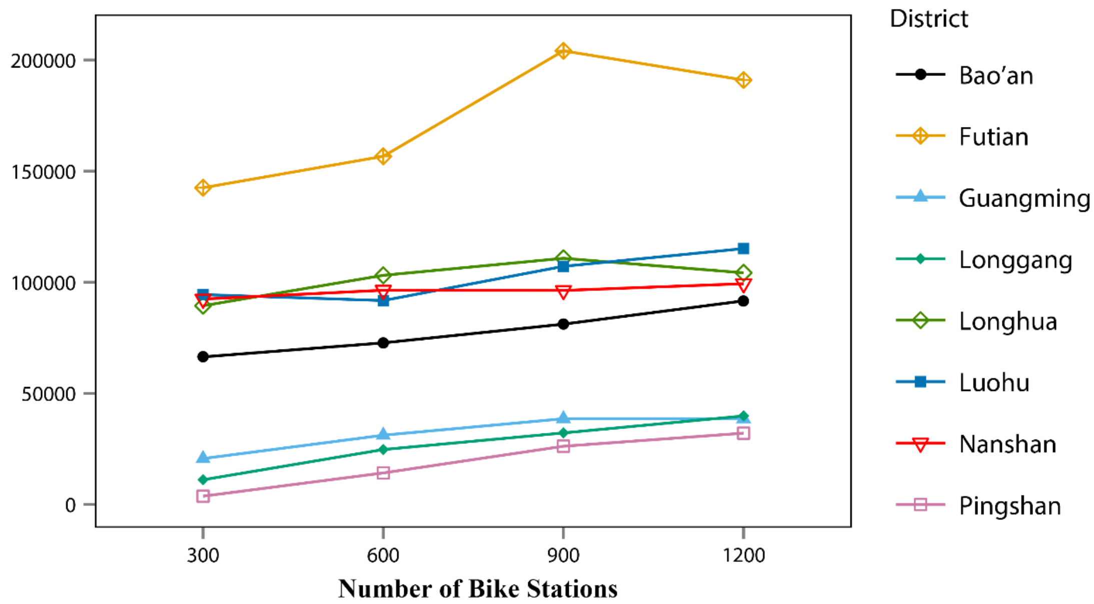



Figure 4.14 Average accessibility of bike stations by districts under different scenarios.

### *4.5.6 Dynamic relationships between incoming and outgoing trips at the bike stations*

In this section, the scenario with $N$ = 1200 is used as an example to illustrate how $netflow_q$ can be used to better understand the relationship between the incoming and outgoing trips at the bike stations. The k-means algorithm is used to group the bike stations into different clusters based on the temporal patterns of $netflow_q$. In order to determine the proper cluster size, we evaluate the changes of total-within cluster variance as we increase the number of clusters. Figure 4.15 shows the total within-cluster variance as we change the cluster size from 1 to 40. It suggests that the variance drops quickly at the beginning of the curve, and then decays slowly when the cluster size becomes larger. In our analysis, we choose 7 as the cluster size to perform the k-means since continuously increasing the number of clusters does not improve the result significantly.

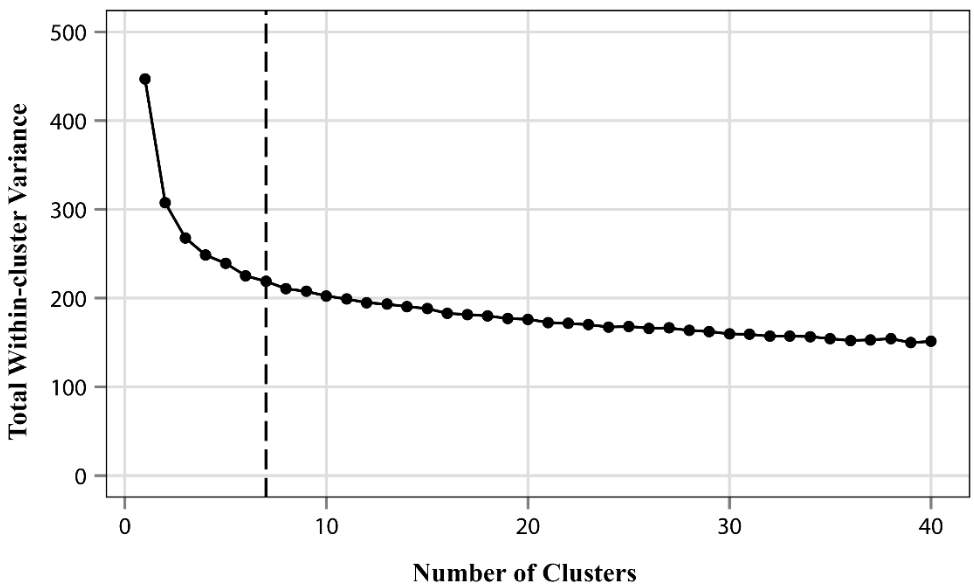



Figure 4.15 Relationship between total within-cluster variance and cluster size in k-means.

Figure 4.16 illustrates the clustering result of $netflow_q$ for the scenario of 1200 bike stations. The average values of $netflow_q$ (i.e., centroids) of the 7 clusters ($C1$ to $C7$) are shown to describe the distinctive characteristics of each cluster type. According to the clustering result, the incoming and outgoing trips of bike stations in $C1$ tended to balance with each other during the entire day. The characteristics of the bike stations in $C1$ can be described as "*mixed*" usage patterns. The bike stations in $C2$ served as trip attractors in the morning, and trip producers in the late afternoon and evening. However, the overall difference between the incoming and outgoing trips was not significant as compared to that of other clusters such as $C3$ and $C4$. Thus, the bike stations in $C2$ can be described as *weak morning attractor – late afternoon and evening producer*. Similarly, bike stations in $C6$ can be described as *weak morning producer – late afternoon and evening attractor*. For $C3$ and $C4$, the values of $netflow_q$ reached almost 0.4 in the morning, which indicates a relatively large difference between the incoming and outgoing

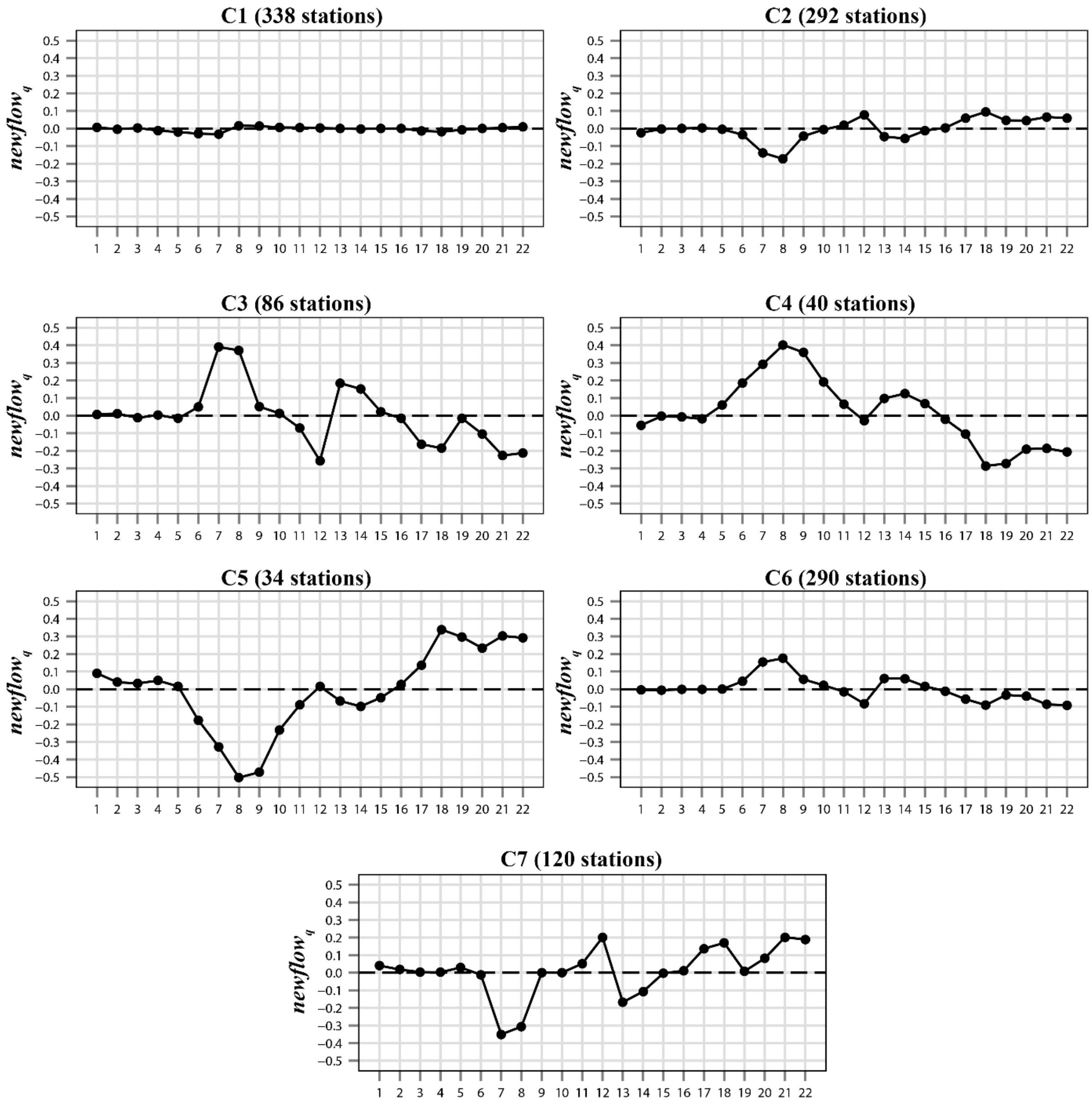



Figure 4.16 Temporal patterns of $netflow_q$ of the 7 clusters derived from k-means.

trips. Hence, the bike stations in $C3$ and $C4$ can be described as *strong morning producer – late afternoon and evening attractor*. The difference between $C3$ and $C4$ is that the morning peak of $C3$ arrived at time interval 7 (i.e., 06:00 – 07:00 to 07:00 – 08:00) and only lasted for two hours. However, the morning peak of $C4$ arrived at time interval 8, and the bike stations in this cluster served as trip producers during the entire morning. Likewise, bike stations in $C5$ and $C7$ can be described as *strong morning attractor – late afternoon and evening producer*. Similar to the difference between $C3$ and $C4$, bike stations in $C5$ served as trip producer during the entire morning, which is different from those in $C7$. We also notice that the bike stations in certain clusters (e.g., $C2$, $C3$, $C6$ and $C7$) tended to have opposite directions of $netflow_q$ at time interval 12 and 13. According to Figure 4.6a, 4.6c and 4.6d, there was a considerable amount of $ND$, $DN$ and $DD$ trip chain segments during noon time. It is likely that certain individuals left their workplaces for particular activities (e.g., dining at restaurants or sleep at home) around this period, and then went back to their workplaces. For example, certain areas with many residential neighborhoods and/or restaurants might gain more incoming trips (than outgoing ones) at time interval 12, but more outgoing trips at time interval 13. On the other hand, certain employment centers might gain more outgoing trips at time interval 12 and vice versa at time interval 13.

Next we further examine the spatial distributions of the seven clusters. As shown in Figure 4.17a, the bike stations in $C1$ are widely dispersed across different districts in Shenzhen. The bike stations are likely to be located in zones with mixed land use patterns. The incoming and outgoing trips at these stations tended to balance each other out during the entire day. $C3$, $C4$ and $C6$ correspond to morning producer – late afternoon and evening attractor. Similar to $C1$, the bike stations in $C6$ are widely distributed across Shenzhen. However, we notice that the bike stations in $C3$ and $C4$ have a general north-south divide. The stations in $C3$ mainly

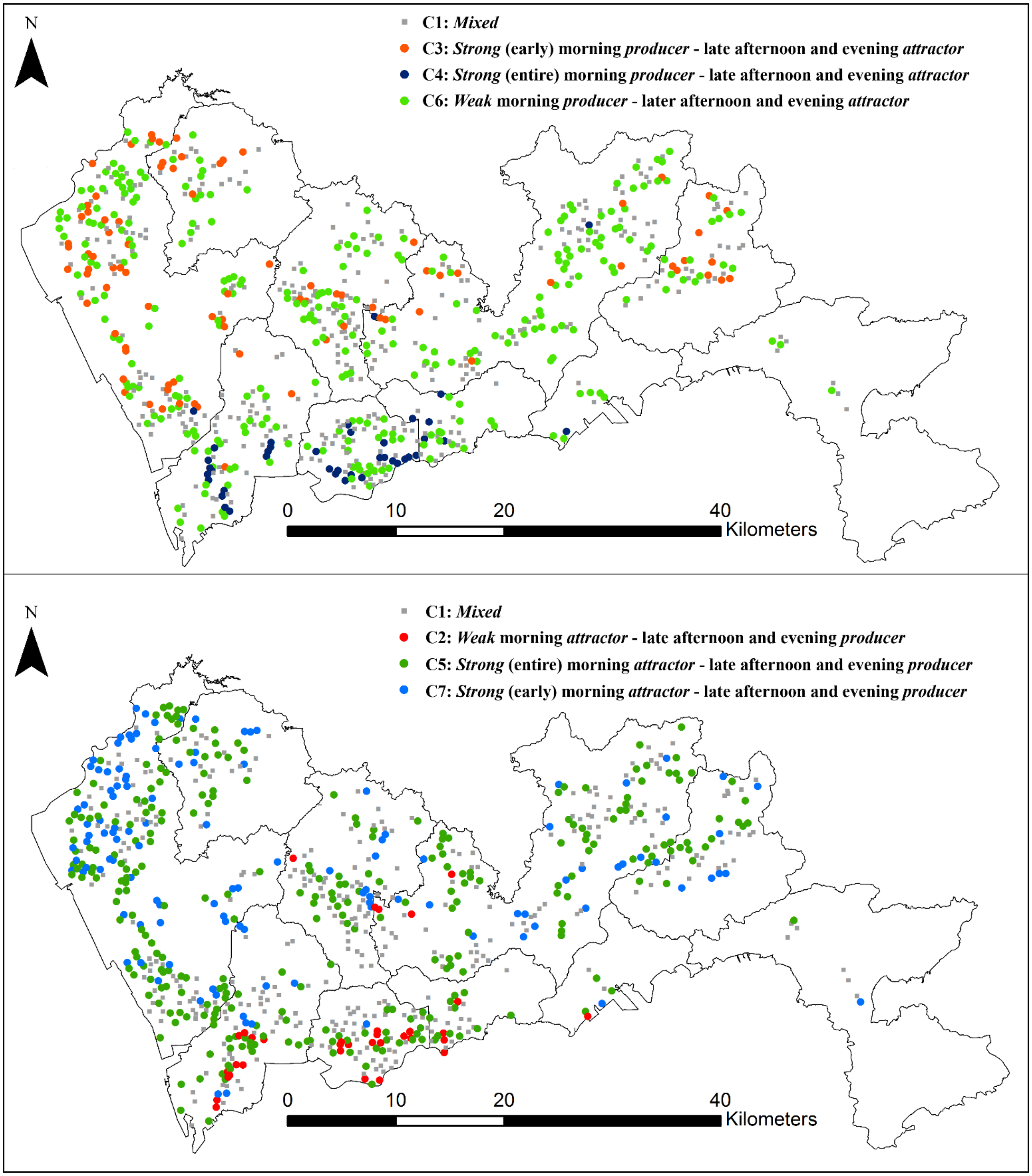



Figure 4.17 (a) Spatial distributions of $C1$, $C3$, $C4$ and $C6$; (b) Spatial distributions of $C1$, $C2$, $C5$ and $C$

distribute in *Guan Wai*, and the ones in $C4$ mainly distribute in *Guan Nei*. As described previously, *Guan Wai* mainly refers to industrial-oriented areas with many migrant workers, while *Guan Nei* offers more diverse employment opportunities related to technology, commerce, education and so forth. The spatial and temporal patterns of $C3$ and $C4$ suggest that people in *Guan Wai* had more rigid work hours than people in *Guan Nei*. Hence, the stations in $C3$ and $C4$ would expect different usage patterns of bicycles. For example, the stations in $C4$ would need enough number of free docks during different time periods in the morning to absorb the incoming trips.

$C2$, $C5$ and $C7$ correspond to morning attractor – late afternoon and evening producer. The bike stations in these clusters (especially $C5$ and $C7$) should have enough free docks in the morning and adequate number of bicycles in the evening. Similarly, there is a north-south divide of the distribution patterns of $C5$ and $C7$. The bike stations in $C5$ cover the major employment and commercial centers in *Guan Nei* (readers could refer to Figure 4.8a for more details). The ones in $C7$ mainly distribute in *Guan Wai*. It requires further investigation (e.g., overlaying the results with land use patterns and POI data) to see if these stations are located around certain factories which serve as the major industry of *Guan Wai*. In general, the difference of people's travel patterns between *Guan Nei* and *Guan Wai* should be regarded as an important factor for the planning and operation of the bike sharing stations in Shenzhen.

## 4.6 Conclusions

A growing number of modern cities are promoting bicycle use to meet the challenges related to public health, traffic congestion, energy consumption and air pollution. Among various considerations of establishing a public bicycle program, knowing where the demands are and when they occur is of primary importance. It is necessary for planners to have useful datasets and effective analytical tools to gain insights of human travel demand that can be potentially served by bicycles. Recent advancements of location-aware technologies have introduced some new data sources which capture the digital footprint of human activities. Such information can be used to better understand the spatio-temporal dynamics of people's short range movements in a city, and further benefit the planning and daily operation of cycling facilities (e.g., bike sharing stations).

Using Shenzhen, China as a case study, this research demonstrates how large-scale cellphone location data can be used to uncover potential demand of bicycle trips in a city, and to provide suggestions on the locations and management strategies of the bike sharing stations. The anchor-point based trajectory segmentation method could effectively partition the individual cellphone trajectories into trip chain segments, which are used to generate potential demand of bicycle trips. By exploring two types of demand (incoming and outgoing trips) derived at the cellphone tower level, our analysis results clearly demonstrate the spatio-temporal dynamics of potential demand and the rhythms of human travel in the city. By applying a maximum coverage location-allocation model, we suggest locations for the bike sharing stations and compare the outcomes among different solutions. We find that although adding equal increment of bike stations ($N$ = 300 to $N$ = 1200 with an increment of 300) would increase the total demand covered, such improvements become less beneficial when $N$ gets large (i.e., 1200). There is a general increase of the overall accessibility of bike stations in Shenzhen as $N$ changes from 300 to 1200. However, the average accessibility of bike stations in districts where potential demands concentrate in a few areas has a diminishing return as $N$ becomes larger. Bike stations in districts where potential demands are more uniform over space have steady improvements as $N$ changes from 300 to 1200. Hence, multiple factors (e.g., total demand covered, accessibility, and investment and operation cost) should be considered when determining the actual number of bike stations to be located.

A k-means algorithm is performed to distinguish the dynamic relationships between the incoming and outgoing trips (e.g., $netflow$) at the bike stations. Seven clusters ($C1$ to $C7$) are derived to illustrate the unique characteristics of them. $C1$ refers to the bike stations where the incoming and outgoing trips balanced with each other during the entire day. $C3$, $C4$ and $C6$ are identified as *morning producer – late afternoon and evening attractor*. The three clusters differ from each other regarding the intensity (e.g., strong or weak) and temporal characteristics (e.g., early morning producer or entire morning producer) of $netflow$. $C2$, $C5$ and $C7$ refer to *morning attractor – late afternoon and evening producer*. The differences among these three clusters are similar to those of $C3$, $C4$ and $C6$. The temporal patterns and spatial distribution of the seven clusters can be used to assist decision making on the distribution and redistribution of bicycles among the bike sharing stations in Shenzhen.

There are several limitations of this study. First, the current research is conducted using a cellphone location dataset on a weekday. Hence, the analysis results mainly reflect the potential demand of bicycle trips on workdays. In the future, it would be beneficial to analyze cellphone location data on weekends and holidays to have a more comprehensive view of travel demand in the city. Second, the suggestions for locating the bike stations are provided from the demand perspective. Other factors such as land topography, safety, current infrastructure of bike lanes, and connectivity to nearby transit stations (Hunt and Abraham 2007; Buehler et al. 2012) should be considered in future studies to evaluate how various cycling facilities and environments affect bicycle usage.

## References

Ahas, R., A. Aasa, Ü. Mark, T. Pae, and A. Kull. 2007. Seasonal tourism spaces in Estonia: Case study with mobile positioning data. *Tourism Management* 28 (3):898-910.

Ahas, R., Aasa, A., Silm, S., Tiru, M., 2010. Daily rhythms of suburban commuters' movements in the Tallinn metropolitan area: case study with mobile positioning data. *Transportation Research Part C: Emerging Technologies* 18, 45-54.

Ahas, R., Silm, S., Järv, O., Saluveer, E., Tiru, M., 2010. Using mobile positioning data to model locations meaningful to users of mobile phones. *Journal of Urban Technology* 17, 3-27.

Alexander, L., S. Jiang, M. Murga, and M. C. González. 2015. Origin–destination trips by purpose and time of day inferred from mobile phone data. *Transportation Research Part C: Emerging Technologies* 58: 240-250.

Barnes, G., Krizek, K., 2005. Estimating bicycling demand. *Transportation Research Record: Journal of the Transportation Research Board*, 45-51.

Bayir, M. A., M. Demirbas, and N. Eagle. 2010. Mobility profiler: A framework for discovering mobility profiles of cell phone users. *Pervasive and Mobile Computing* 6 (4):435-454.

Becker, R., Cáceres, R., Hanson, K., Isaacman, S., Loh, J.M., Martonosi, M., Rowland, J., Urbanek, S., Varshavsky, A., Volinsky, C., 2013. Human mobility characterization from cellular network data. *Communications of the ACM* 56, 74-82.

Buehler, R. 2012. Determinants of bicycle commuting in the Washington, DC region: The role of bicycle parking, cyclist showers, and free car parking at work. *Transportation Research Part D: Transport and Environment* 17 (7):525-531.

Calabrese, F., M. Diao, G. Di Lorenzo, J. Ferreira, and C. Ratti. 2013. Understanding individual mobility patterns from urban sensing data: A mobile phone trace example. *Transportation Research Part C: Emerging Technologies* 26:301-313.

Candia, J., González, M.C., Wang, P., Schoenharl, T., Madey, G., Barabási, A.-L., 2008. Uncovering individual and collective human dynamics from mobile phone records. *Journal of Physics A: Mathematical and Theoretical* 41, 224015.

Cervero, R., Duncan, M., 2003. Walking, bicycling, and urban landscapes: evidence from the San Francisco Bay Area. *American Journal of Public Health* 93, 1478-1483.

China to create largest mega city in the world with 42 million people. www. telegraph.co.uk. Last accessed on 2015-07-18.

Cho, E., S. A. Myers, and J. Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining. pp. 1082-1090.

Church, R., and C. R. Velle. 1974. The maximal covering location problem. *Papers in Regional Science* 32 (1):101-118.

Clark, D. 1997. Estimating future bicycle and pedestrian trips from a travel demand forecasting model. Proceedings of the 67th Annual Meeting of the Institute of Transportation Engineers. Washington DC, USA.

Csáji, B.C., Browet, A., Traag, V.A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., Blondel, V.D., 2013. Exploring the mobility of mobile phone users. *Physica A: Statistical Mechanics and its Applications* 392, 1459-1473.

DeMaio, P.J., 2003. Smart bikes: Public transportation for the 21st century. *Transportation Quarterly* 57, 9-11.

DeMaio, P., and J. Gifford. 2004. Will smart bikes succeed as public transportation in the United States? *Journal of Public Transportation* 7 (2):1-15.

DeMaio, P., 2009. Bike-sharing: History, impacts, models of provision, and future. *Journal of Public Transportation* 12, 3.

de Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports* 3(1376).

Dijst, M., 1999. Two-earner families and their action spaces: A case study of two Dutch communities. *GeoJournal* 48, 195-206.

Dong, H., M. Wu, X. Ding, L. Chu, L. Jia, Y. Qin, and X. Zhou. 2015. Traffic zone division based on big data from mobile phone base stations. *Transportation Research Part C: Emerging Technologies* 58: 278-291.

Wardman, M., M. Tight, and M. Page. 2007. Factors influencing the propensity to cycle to work. *Transportation Research Part A: Policy and Practice* 41 (4):339-350.

García-Palomares, J.C., Gutiérrez, J., Latorre, M., 2012. Optimizing the location of stations in bike-sharing programs: a GIS approach. *Applied Geography* 35, 235-246.

Golob, T.F., Hensher, D.A., 2007. The trip chaining activity of Sydney residents: a cross-section assessment by age group with a focus on seniors. *Journal of Transport Geography* 15(4): 298-312.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196):779-782.

Hägerstraand, T. 1970. What about people in regional science? *Papers in Regional Science* 24 (1):7-24.

Hakimi, S. L. 1965. Optimum distribution of switching centers in a communication network and some related graph theoretic problems. *Operations Research* 13 (3):462-475.

Hansen, W.G., 1959. How accessibility shapes land use. *Journal of the American Institute of Planners* 25(2) 73-76.

Harkey, D., D. Reinfurt, and M. Knuiman. 1998. Development of the bicycle compatibility index. *Transportation Research Record: Journal of the Transportation Research Board* (1636):13-20.

Hunt, J.D., Abraham, J., 2007. Influences on bicycle use. *Transportation* 34, 453-470.

Iacono, M., Krizek, K.J., El-Geneidy, A., 2010. Measuring non-motorized accessibility: issues, alternatives, and execution. *Journal of Transport Geography* 18, 133-140.

Institute for Transportation & Development Policy. The Bike-share Planning Guide. https://www.itdp.org/wp-content/uploads/2014/07/ITDP_Bike_Share_Planning_Guide.pdf. (last accessed 25 October 2015).

Iqbal, M. S., C. F. Choudhury, P. Wang, and M. C. González. 2014. Development of origin–destination matrices using mobile phone call data. Transportation Research Part C: Emerging Technologies 40:63-74.

Isaacman, S., Becker, R., Cáceres, R., Martonosi, M., Rowland, J., Varshavsky, A., Willinger, W., 2012. Human mobility modeling at metropolitan scales, Proceedings of the 10th international conference on Mobile systems, applications, and services. ACM, pp. 239-252.

Landis, B.W., 1996. Bicycle system performance measures. *ITE journal* 66, 18-26.

Larsen, J., Patterson, Z., El-Geneidy, A., 2013. Build it. But where? The use of geographic information systems in identifying locations for new cycling infrastructure. International *Journal of Sustainable Transportation* 7, 299-317.

List of administrative divisions of Shenzhen. 2015. In *Wikipedia*. https://en.wikipedia.org/wiki/List_of_administrative_divisions_of_Shenzhen (last accessed 03 October 2015).

Long, Y., Y. Zhang, and C. Cui. 2012. Identifying commuting pattern of Beijing using bus smart card data. *Acta Geographica Sinica* 67 (10):1339-1352.

Martinez, L. M., L. Caetano, T. Eiró, and F. Cruz. 2012. An optimisation algorithm to establish the location of stations of a mixed fleet biking system: an application to the city of Lisbon. *Procedia-Social and Behavioral Sciences* 54:513-524.

McGuckin, N., Zmud, J., Nakamoto, Y., 2005. Trip-chaining trends in the United States: understanding travel behavior for policy making. *Transportation Research Record: Journal of the Transportation Research Board* 1917:199-204.

Ning, X., Ling, Y., Hu, J., 2014. Identifying Home-Work Locations from Short-term, Large-scale, and Regularly Sampled Mobile Phone Tracking Data. Geomatics and Information Science of Wuhan University 39(6): 750-756.

Phithakkitnukoon, S., T. Horanont, G. Di Lorenzo, R. Shibasaki, and C. Ratti. 2010. Activity-aware map: Identifying human daily activity pattern using mobile phone data. In *Human Behavior Understanding* 6219:14-25.

Porter, C., Suhrbier, J., Schwartz, W., 1999. Forecasting bicycle and pedestrian travel: state of the practice and research needs. *Transportation Research Record: Journal of the Transportation Research Board*, 94-101.

Ratti, C., Sevtsuk, A., Huang, S., Pailer, R., 2007. Mobile landscapes: Graz in real time. Springer.

Reades, J., Calabrese, F., Ratti, C., 2009. Eigenplaces: analysing cities using the space-time structure of the mobile phone network. *Environment and Planning B: Planning and Design* 36, 824-836.

Rushton, G. 1979. Optimal location of facilities: COM Press Wentworth, New Hampshire.

Rybarczyk, G., Wu, C., 2010. Bicycle facility planning using GIS and multi-criteria decision analysis. *Applied Geography* 30, 282-293.

Schönfelder, S., Axhausen, K.W., 2003. Activity spaces: measures of social exclusion? *Transport policy* 10, 273-286.

Shaheen, S., Guzman, S., Zhang, H., 2010. Bikesharing in Europe, the Americas, and Asia: past, present, and future. *Transportation Research Record: Journal of the Transportation Research Board*, 159-167.

Shenzhen Daily. 2012. "Shenzhen: Most crowded in China". http://szdaily.sznews.com/html/2012-05/30/content_2063502.htm (last accessed 03 October 2015).

Shenzhen Special Economic Zone. 2015. In *Wikipedia*. https://en.wikipedia.org/wiki/Shenzhen_Special_Economic_Zone (last accessed on 3 October 2015).

Strathman, J.G., Dueker, K.J., Davis, J.S., 1994. Effects of household structure and selected travel characteristics on trip chaining. *Transportation* 21(1) 23-45.

Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010. Limits of predictability in human mobility. *Science* 327, 1018-1021.

Toregas, C., R. Swain, C. ReVelle, and L. Bergman. 1971. The location of emergency service facilities. *Operations Research* 19 (6):1363-1373.

Transport Commission of Shenzhen Municipality. 2011. 深圳市步行和自行车交通规划及设计导则 (Guidelines of transport planning and design for pedestrian and bicycle systems in Shenzhen). http://www.szpl.gov.cn/xxgk/ztzl/zxcgh/jtghcgg.pdf (last accessed 31 August 2015).

Vieira, M. R., V. Frias-Martinez, N. Oliver, and E. Frias-Martinez. 2010. Characterizing dense urban areas from mobile phone-call data: Discovery and social dynamics. Social 2010 IEEE Second International Conference on Computing (SocialCom).

Wang, Z.-g., Kong, Z., Xie, J.-h., YIN, L.-e., 2009. The 3rd generation of bike sharing systems in Europe: programs and implications. *Urban Transport of China* 4, 7-12.

Wang, P., T. Hunter, A. M. Bayen, K. Schechtner, and M. C. González. 2012. Understanding road usage patterns in urban areas. *Scientific reports* 2:1001.

Wei, L., Yan, X., 2005. Transformation Of "Urban Village" And Feasible Mode [J]. *City Planning Review* 7.

Yuan, Y., M. Raubal, and Y. Liu. 2012. Correlating mobile phone usage and travel behavior–A case study of Harbin, China. *Computers, Environment and Urban Systems* 36 (2):118-130.

# CHAPTER 5

# CONCLUSION

## 5.1 Summary of Dissertation Research

Understanding how people move around in their daily lives has always been an important research topic in geography. The emergence of large individual tracking datasets has brought new opportunities to the understanding of human travel behavior and daily activity patterns in urban environment. Due to the large spatial, temporal and population coverage, mobile phone location data enables studies of human activity spaces and mobility patterns at an unprecedented scale. Although mobile phone location data provides many new opportunities for human behavior research, challenges still present in many perspectives. Firstly, traditional travel surveys are usually designed to answer specific research questions on human travel behavior. The information collected by these travel surveys (e.g., individual socioeconomic and demographic characteristics) is not available in mobile phone location data, which is usually collected by phone companies for business operation purposes. Secondly, mobile phone location data include limited information which suggests the specific locations visited by the individuals. Such datasets reflect a partial aspects of individual daily activity patterns. Furthermore, data noise and spatial uncertainty often exist in mobile phone location data for particular reasons (e.g., data collection mechanism and load balancing of cellphone signal). Hence, we need new methods to process and analyze mobile phone location data to uncover useful knowledge of human activity spaces and mobility patterns.

This study proposes several approaches for processing and analyzing two types of mobile phone location data (CDRs and Actively Tracked Mobile Phone Location Data) as part of a broad effort of understanding human mobility patterns and people's use of space in geographic research. The main purpose is to demonstrate how large scale mobile phone location data can be leveraged to uncover important characteristics of individual activity space, and how such information could be used to assist urban and transportation planning. This dissertation research addresses three particular research questions. The major findings and contributions are summarized as follows.

Chapter 2 introduces a home-based approach to understanding human activity space using a CDR dataset in Shenzhen, China. Different from previous activity space measures which consider the arithmetic mean of individual daily activities as the center of activity space (Gonzalez, et al. 2008; Song et al. 2010; Yuan, Raubal and Liu 2012), this approach considers home location as an important anchor point and analyzes how people's daily activities take place around this anchor point. There are several advantages of adopting this approach. First, the modified standard distance introduced in the study provides a measure of activity space size that can be used to evaluate how individuals use urban space around where they live. Second, individuals can be grouped in space based on the estimated home anchor points to derive aggregate activity space patterns. The aggregate activity space patterns reflect group behavioral characteristics at different places in a city. As the distinctive characteristics of activity space patterns are potentially tied to the socioeconomic and demographic characteristics of the built environment, such information can be used to generate reasonable hypotheses of human travel behavior. For example, the general "north-south divide" of Shenzhen identified in this dissertation can be further used to evaluate how the regional differences of economic and employment structures affect people's activity space. The study also yields some insights into important societal issues, such as the small activity space sizes of people who live in the "urban villages" (Wei and Yan 2005). Moreover, the aggregate activity space patterns derived at the mobile phone tower level can be further integrated with other datasets which include important explanatory variables for travel behavior studies and policy analysis.

Chapter 3 introduces an analytical framework that is capable of examining multiple determinants of individual activity space simultaneously using mobile phone location data. As pointed out by Golldge and Stimson (1997), the spatial extent, frequent locations and movements are three major determinants of individual activity space. These determinants provide much needed information for the understanding of human mobility patterns. Due to the large data volume (i.e. sample sizes) to be analyzed in mobile phone location data, it becomes difficult to summarize the major characteristics as well as the interactions of these determinants. To fill the research gap, this study develops three intuitive mobility indicators, which are daily activity range, number of activity anchor points, and frequency of movements to represent three major determinants of individual activity space. These mobility indicators are easy to understand and can be combined to describe the major characteristics of an individual's activity space. The association rules in data mining research (Han, Kember and Pei 2011) are applied to examine how the three mobility indicators are related to each other in an individual's activity space. Using actively tracked mobile phone location data collected in two major cities in China, the study demonstrates how the analytical framework can be used to summarize and compare human activity spaces systematically across different population groups or geographic regions. The proposed methods can be used to derive various "signatures" of people's daily activity patterns. These signatures serve as critical information for understanding the basic rules of how people move around in their daily lives. Moreover, the analytical methods (e.g., association rules) are scalable over very large datasets, which means that the framework can be easily extended for studies that involve: (1) many activity space determinants; (2) multiple geographic regions (e.g., cities), and (3) very large mobile phone location data or similar individual tracking datasets. Overall, this study serves as a bridge which connects traditional activity space concepts with big data analytics to answer critical questions of individual activity space (i.e., how far, how many and how frequent).

Chapter 4 investigates human mobility patterns from another perspective which emphasizes on travel demand estimation and transportation planning. In particular, the study uses an actively tracked mobile phone location dataset in Shenzhen, China to uncover potential demand of bicycle trips in the city. Several contributions of this study are worth noting. First, the anchor-point based trajectory segmentation method could effectively partition cellphone trajectories into four types of trip chain segments. These trip chain segments reflect human movements around or among important individual activity anchor points (i.e., night-time and day-time anchor points). These activity anchor points are important activity origins/destinations when individuals plan their daily trips with single or multiple purposes. Thus, the derived segments serve as the fundamental elements to understand individual travel behavior using mobile phone location data. Second, the two indicators ($inflow_p$ and $outflow_p$) derived at the cellphone tower level clearly reflect the dynamics of human travel demand at different places in the city and time of a day. By suggesting the locations of bike sharing stations using a location-allocation model, the study evaluates the dynamic relationships between the incoming and outgoing trips at the suggested bike stations in the city. The seven clusters derived in the study clearly reflect the regional differences and rhythms of human travel patterns. Such information is useful to the planning and daily operation of cycling facilities in a city. Although the current study focuses on the potential demand of bicycle trips, the proposed methods can be applied to understand travel demand related to other transportation modes by adjusting the filtering strategy of trip chain segments introduced in the study. For example, the same mobile phone location dataset can be used to provide suggestions on where to build pedestrian walkways to accommodate human movements that were very short in distance. More importantly, the derived trip chain segments can be further analyzed to estimate different types of individual daily trips (e.g., home-based work trips, home-based other trips and non-home based trips). Such information can be used along with travel surveys to improve or calibrate existing travel demand models (e.g., four-step model and activity-based models) to assist a variety of applications in urban and transportation planning.

## 5.2 Future Research Directions

This dissertation research serves as the starting point of investigating human mobility patterns and activity spaces using large scale mobile phone location data. According to the current analysis results and findings, several things need to be improved in the future. First, it would be useful to evaluate and compare different measures that describe important characteristics of individual activity space under specific scenarios. The modified standard distance and daily activity range introduced in this research are two particular ways to represent individual activity space. There are many other activity space measures such as radius of gyration (Song et al. 2010), confidence ellipse (Schönfelder and Axhausen 2004), standard distance (Buliung and Kanaroglou 2006) and standard deviational ellipse (Lefever 1926). As mobile phone location data (e.g., CDRs) can be sparse in both space and time, these measures could produce different results of individual activity space in terms of spatial extent, direction and shape. Thus, it would be helpful to evaluate their differences and similarities using mobile phone location data, and discuss their suitability of representing individual activity space under particular research contexts (Sherman et al. 2005). Second, the association rules used in Chapter 3 uncover many "signatures" of human activity space patterns in Shenzhen and Shanghai, China. However, it requires additional efforts to investigate potential driving forces of these activity space patterns. For example, what are potential causes of the “immobility” of individuals who stayed around a particular location during the entire day in these two cities? How do urban forms and neighborhood designs affect people’s use of space in Shenzhen and Shanghai? Answering these questions would advance our understanding of human activity spaces and their relationships with the built environment.

This dissertation research also brings many new opportunities and perspectives for future studies. First, the geographic disparities of human mobility patterns and their use of urban space have been discussed from different perspectives in the above three chapters. The analysis results demonstrate that mobile phone location data can help us understand the land use patterns and socioeconomic characteristics of the built environment (Soto et al. 2011; Soto and Frías-Martínez 2011; Toole et al. 2012; Pei et al. 2014). The current analyses of human activity space can be extended to follow-up studies which focus on questions related to social segregation (Silm and Ahas 2014), job-housing balance (Cervero 1989; Cervero 1996; Wang and Chai 2009), urban functional regions (Liu et al. 2012; Yuan, Zheng and Xie 2012) and so forth. However, mobile phone location data provide limited information for answering these questions. It is thus important to fuse information from multiple data sources to gain useful knowledge of human mobility patterns (Zheng 2015). For example, it would be useful to incorporate land use and socio-demographic characteristics, point of interest (POI), and even travel surveys into the analysis to better explain the identified patterns of human activity space.

Second, many previous studies aim at examining human mobility patterns from an aggregate perspective without looking into each individual’s characteristics. The information embedded in mobile phone location data inspires us to move from an aggregate approach towards a disaggregate approach (Shaw, Yu and Bombom 2008; Shaw, 2009). There is a need to “move beyond place-based perspective to include a people-based perspective that focuses on individuals in space and time and their allocation of activities in physical and virtual worlds” (Miller 2007, p. 527). For example, the four types of trip chain segments derived in Chapter 4 enable us to examine each individual’s movement patterns as a function of location and time. More efforts should be made in the future to analyze longitudinal mobile phone location data to uncover some basic laws that govern how individuals move around in their daily lives. Such information serves as an important element to modeling human mobility patterns (Rhee et al. 2011; Isaacman et al. 2012). Large scale human mobility models and simulations can be developed and validated using mobile phone location data. This will benefit applications across several domains such as emergency response, epidemiology, location-based services, and travel demand forecast.

Last but not least, the rapid development of information and communication technologies (ICT) has brought remarkable changes to our society. The technological advancements, such as the proliferation of e-commerce and online social networking tools are changing the ways people interact with the physical and virtual world (Shaw and Yu 2009). By analyzing large scale mobile phone location data, some studies have demonstrated that the similarities between individuals’ movements are correlated to their proximity in social networks (Cho, Myers and Leskovec 2011; Wang et al. 2011). Thus, it is important to study human mobility patterns and their use of physical space by considering their activities and interactions with other people in the virtual space. For example, how can we represent human activity space in their social networks? Will human mobility patterns be affected by their activity levels in the virtual world? These are the research questions waiting to be answered in the future.

It is necessary to note that privacy should be protected in studies using mobile phone location data. There have been some studies that investigate privacy related issues when large individual tracking datasets are used in human behavioral research (Abul, Bonchi and Nanni 2008; de Montjoye et al. 2014). It is important to preserve individual privacy as a premise of research practice, and use knowledge extracted from mobile phone location data to advance our understanding of human behavior, and benefit urban and transportation planning.

## References


Abul, O., F. Bonchi, and M. Nanni. 2008. Never walk alone: Uncertainty for anonymity in moving objects databases. In Proceedings of IEEE 24th International Conference on Data Engineering pp. 376-385.

Buliung, R. N., and P. S. Kanaroglou. 2006. A GIS toolkit for exploring geographies of household activity/travel behavior. *Journal of Transport Geography* 14 (1):35-51.

Cervero, R. 1989. Jobs-housing balancing and regional mobility. *Journal of the American Planning Association* 55 (2):136-150.

Cervero, R. 1996. Jobs-housing balance revisited: trends and impacts in the San Francisco Bay Area. *Journal of the American Planning Association* 62 (4):492-511.

Cho, E., S. A. Myers, and J. Leskovec. 2011. Friendship and mobility: user movement in location-based social networks. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining pp. 1082-1090.

de Montjoye, Y.-A., C. A. Hidalgo, M. Verleysen, and V. D. Blondel. 2013. Unique in the Crowd: The privacy bounds of human mobility. *Scientific reports* 3.

Golledge, R., and R. Stimson. 1997. *Spatial Behavior*. Guilford, London.

Gonzalez, M. C., C. A. Hidalgo, and A.-L. Barabasi. 2008. Understanding individual human mobility patterns. *Nature* 453 (7196):779-782.

Han, J., Kamber, M., Pei, J. (2011). *Data mining: concepts and techniques, 3rd ed*. Elsevier: Morgan Kaufmann.

Isaacman, S., R. Becker, R. Cáceres, M. Martonosi, J. Rowland, A. Varshavsky, and W. Willinger. 2012. Human mobility modeling at metropolitan scales. In Proceedings of the 10th international conference on Mobile systems, applications, and services. ACM, pp. 239-252.

Lefever, D. W. 1926. Measuring geographic concentration by means of the standard deviational ellipse. *American Journal of Sociology* 32(1):88-94.

Liu, Y., F. Wang, Y. Xiao, and S. Gao. 2012. Urban land uses and traffic 'source-sink areas': Evidence from GPS-enabled taxi data in Shanghai. *Landscape and Urban Planning* 106 (1):73-87.

Miller, H. 2007. Place-based versus people-based geographic information science. *Geography Compass* 1 (3):503-535.

Pei, T., S. Sobolevsky, C. Ratti, S.-L. Shaw, T. Li, and C. Zhou. 2014. A new insight into land use classification based on aggregated mobile phone data. *International Journal of Geographical Information Science* 28(9): 1988-2007.

Rhee, I., M. Shin, S. Hong, K. Lee, S. J. Kim, and S. Chong. 2011. On the levy-walk nature of human mobility. *IEEE/ACM Transactions on Networking (TON)* 19 (3):630-643.

Schönfelder, S., Axhausen, K.W., 2003. Activity spaces: measures of social exclusion? *Transport policy* 10, 273-286.

Schönfelder, S., and K. Axhausen. 2004. Structure and innovation of human activity spaces. *Arbeitsberichte Verkehrs-und Raumplanung* 258:1-40.

Shaw, S-L. 2009. Individual-based Tracking Data: Potentials and Challenges to Transportation Geography. Fleming Lecture of 2009 AAG Meeting. Las Vegas, NV.

Shaw, S. L., H. Yu, and L. S. Bombom. 2008. A space-time GIS approach to exploring large individual-based spatiotemporal datasets. *Transactions in GIS* 12 (4):425-441.

Shaw, S.-L., and H. Yu. 2009. A GIS-based time-geographic approach of studying individual activities and interactions in a hybrid physical–virtual space. *Journal of Transport Geography* 17 (2):141-149.

Sherman, J. E., J. Spencer, J. S. Preisser, W. M. Gesler, and T. A. Arcury. 2005. A suite of methods for representing activity space in a healthcare accessibility study. *International Journal of Health Geographics* 4 (1):24.

Silm, S. and R. Ahas. 2014. Ethnic Differences in Activity Spaces: A Study of Out of-Home Nonemployment Activities with Mobile Phone Data. *Annals of the Association of American Geographers* 104(3):542-559.

Soto, V., and E. Frías-Martínez. 2011. Automated land use identification using cell-phone records. In Proceedings of the 3rd ACM international workshop on MobiArch, pp. 17-22.

Soto, V., V. Frias-Martinez, J. Virseda, and E. Frias-Martinez. 2011. Prediction of socioeconomic levels using cell phone records. In User Modeling, Adaption and Personalization, 377-388: Springer.

Song, C., Qu, Z., Blumm, N., Barabási, A.-L., 2010. Limits of predictability in human mobility. *Science* 327, 1018-1021.

Toole, J. L., M. Ulm, M. C. González, and D. Bauer. 2012. Inferring land use from mobile phone activity. In Proceedings of the ACM SIGKDD international workshop on urban computing. ACM, pp. 1–8.

Wang, D., and Y. Chai. 2009. The jobs–housing relationship and commuting in Beijing, China: the legacy of Danwei. *Journal of Transport Geography* 17 (1):30-38.

Wang, D., D. Pedreschi, C. Song, F. Giannotti, and A.-L. Barabasi. 2011. Human mobility, social ties, and link prediction. In Proceedings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.1100-1108.

Wei, L., Yan, X., 2005. Transformation of "Urban Village" And Feasible Mode [J]. *City Planning Review* 7.

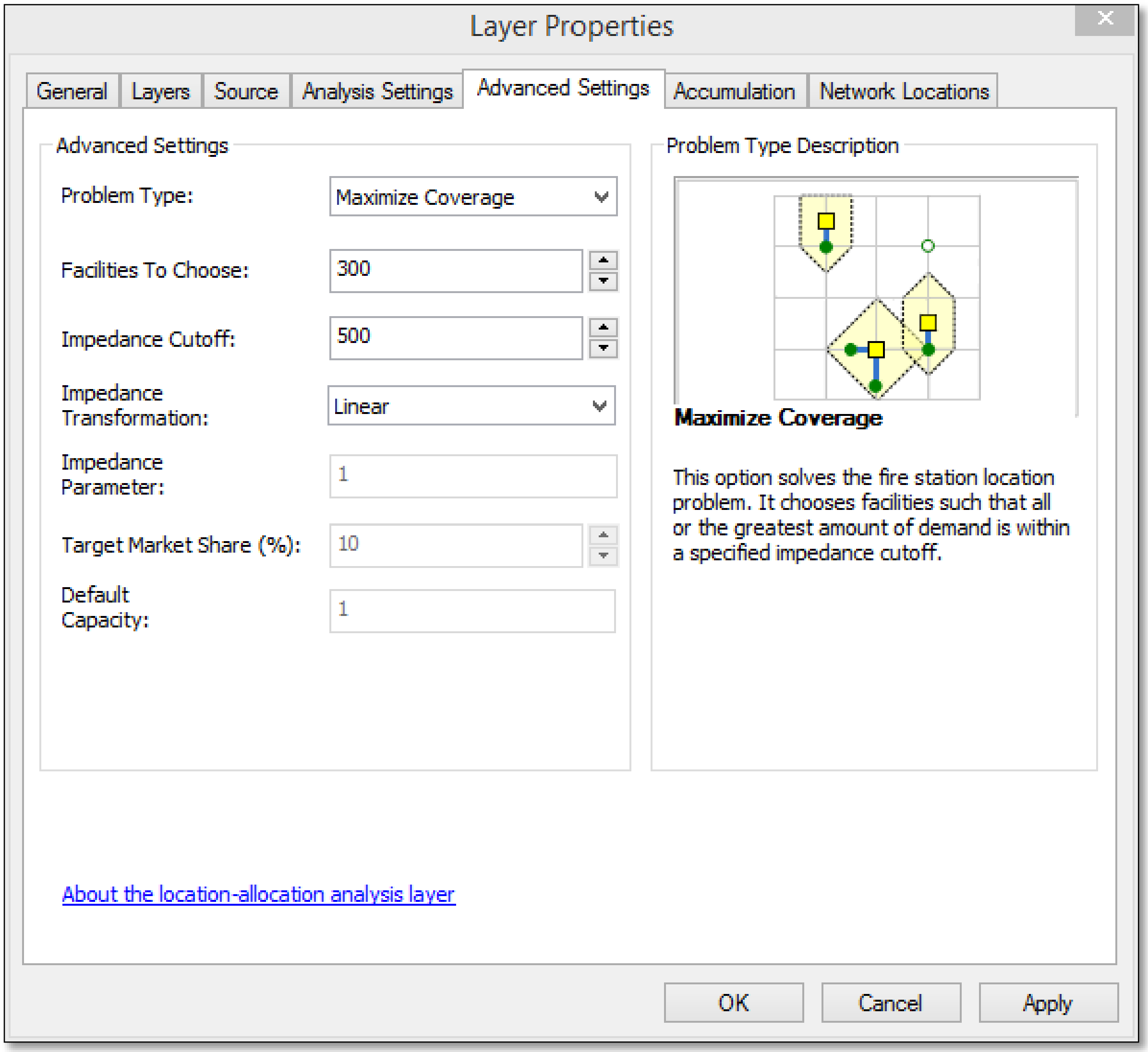
Yuan, Y., M. Raubal, and Y. Liu. 2012. Correlating mobile phone usage and travel behavior–A case study of Harbin, China. *Computers, Environment and Urban Systems* 36 (2):118-130.

Yuan, J., Y. Zheng, and X. Xie. 2012. Discovering regions of different functions in a city using human mobility and POIs. In Proceedings of the 18th ACM SIGKDD international conference on Knowledge discovery and data mining, pp.186-194.

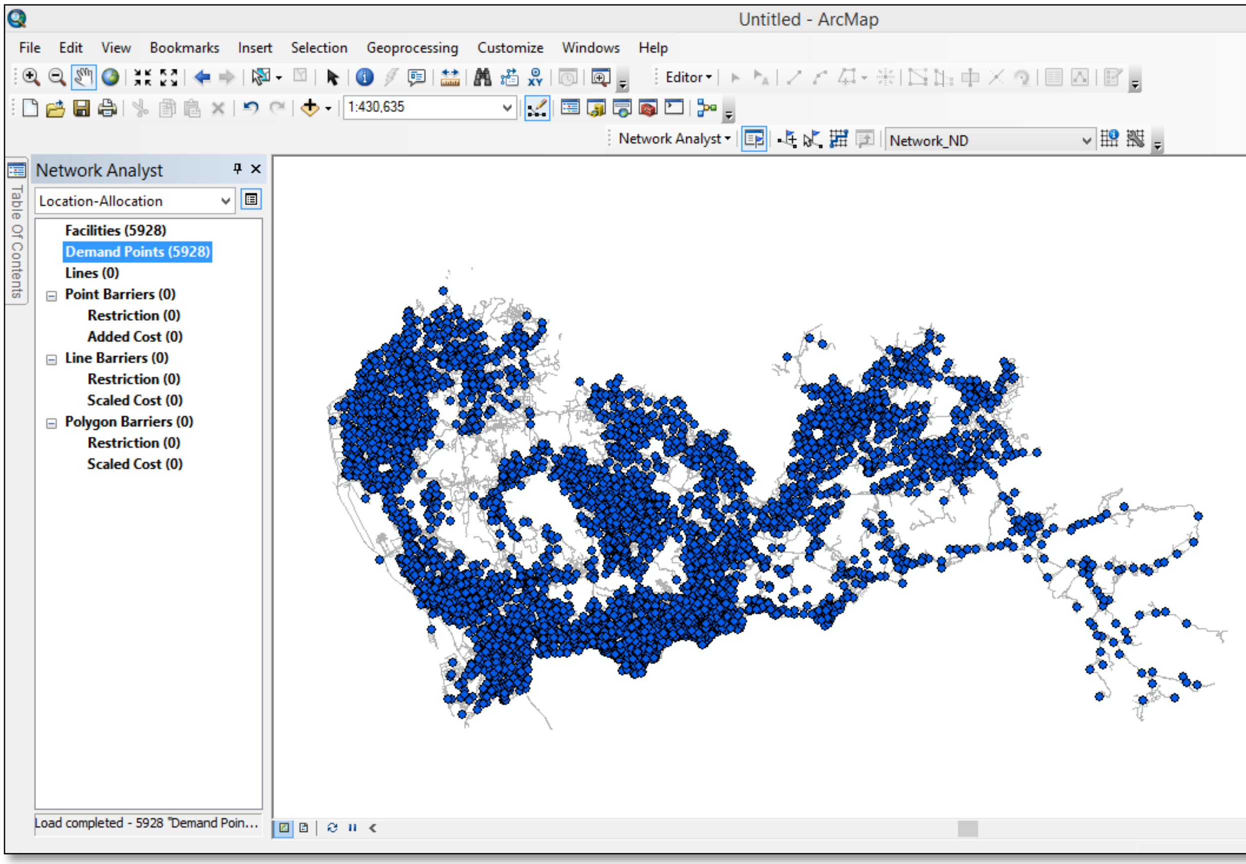
Zheng, Y. 2015. Methodologies for Cross-Domain Data Fusion: An Overview. *IEEE Transactions On Big Data*. pp. 1-18.

# APPENDIX

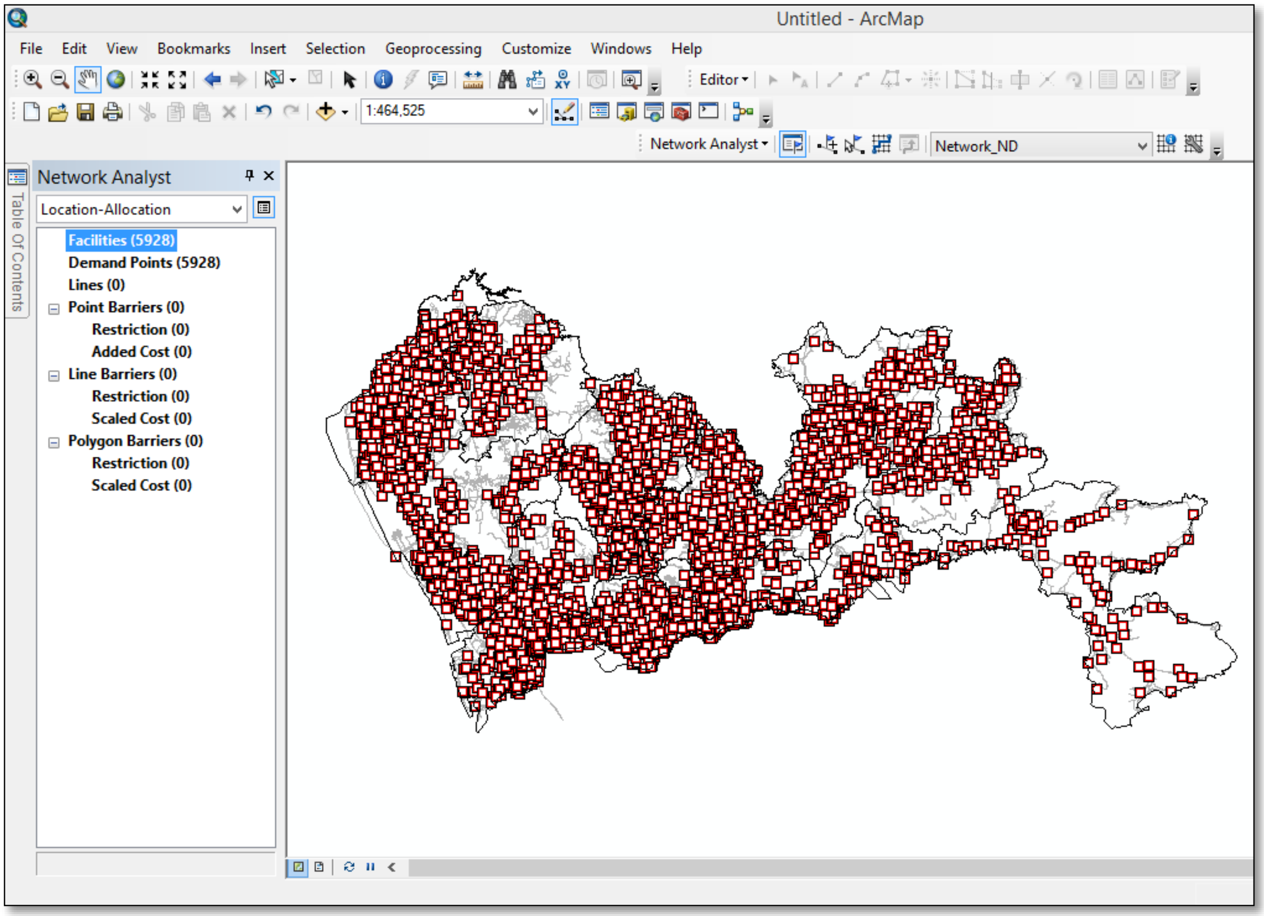
**Appendix 1. The *maximum coverage* module in the Network Analyst toolbox in ArcGIS 10.1.**



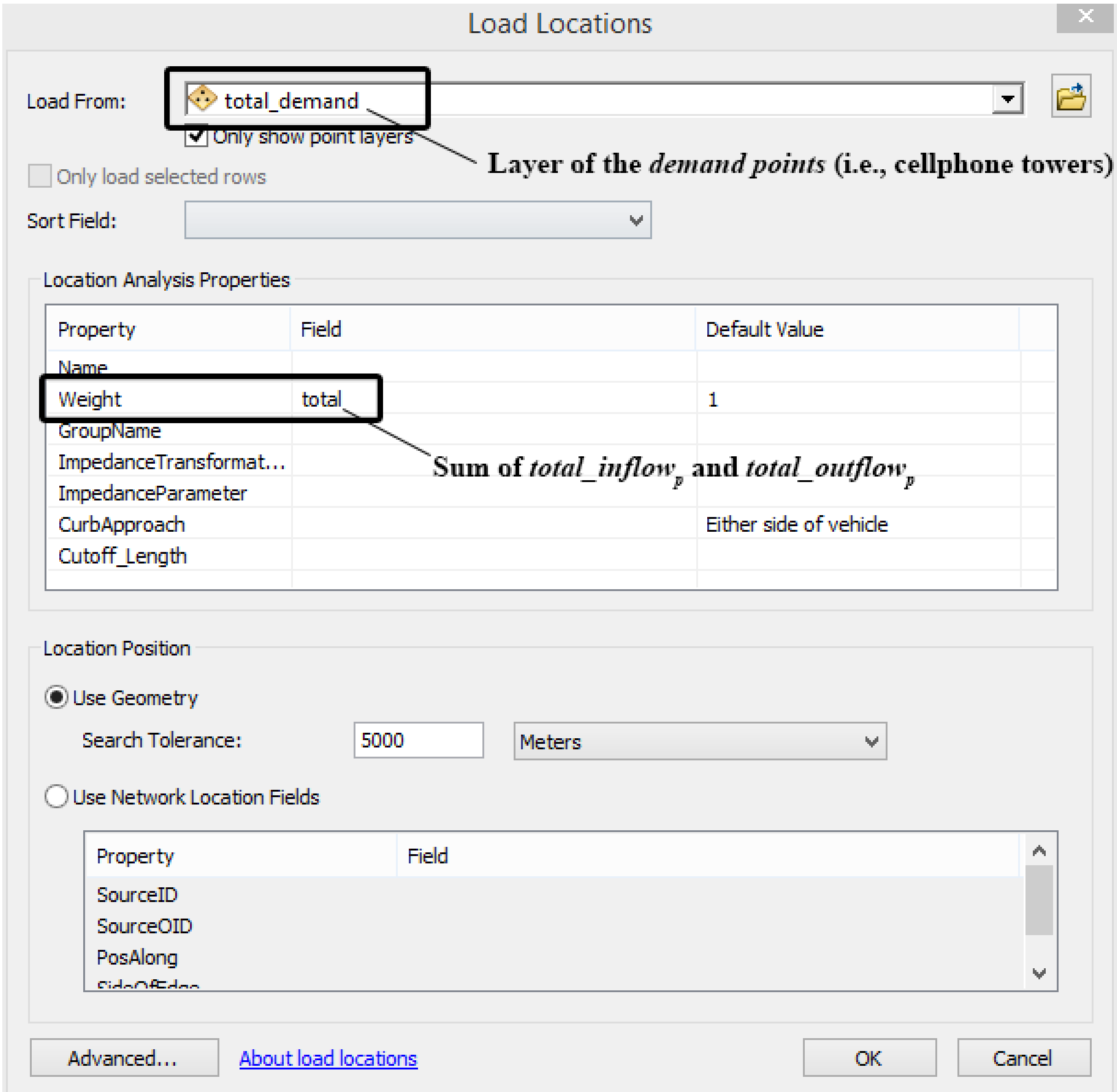
**Appendix 2. The 5928 *demand points* (i.e. blue circles in the map) defined in the *maximum coverage* module.**



**Appendix 3. The 5928 *candidate locations of facilities* (i.e. squares in the map) defined in the *maximum coverage* module.**



**Appendix 4. The *weight* at each demand point $p$ (i.e., cellphone tower $p$) is set as the sum of $total\_inflow_p$ and $total\_outflow_p$.**



# VITA

Yang Xu was born and raised in Wuhan, the People's Republic of China. After finishing high school, he entered Wuhan University in 2005. He received a Bachelor of Sciences degree in Remote Sensing and Photogrammetry in 2009, and a Master of Sciences degree in Geographic Information System (GIS) in 2011. In the summer of 2011, he came to the U.S. and entered the geography department at the University of Tennessee, Knoxville, where he completed his Doctoral of Philosophy degree in December 2015. He has been active in publishing papers in peer-reviewed journals (e.g., *Transportation, Annals of the Association of American Geographers,* and *Environmental Modeling & Software*) and delivering presentations at professional conferences. His research interests include geographic information sciences, transportation, spatial-temporal data mining, and information visualization.